Evaluating pause particles and their functions in natural and synthesized speech in laboratory and lecture settings

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von

Mikey Elmers
geboren in Spokane

Saarbrücken, im August 2023

Dekanin der Fakultät P: Univ.-Prof. Dr. Stefanie Haberzettl

Erstberichterstatter: Prof. Dr. Bernd Möbius

Zweitberichterstatter: Prof. Dr. Joakim Gustafsson

Tag der letzten Prüfungsleistung: November 15, 2023

Acknowledgements

I would like to take this opportunity to express my deep gratitude and appreciation to everyone who supported and contributed to the completion of this dissertation. Without their invaluable assistance and encouragement, this work would not have been possible.

First and foremost, I extend my heartfelt thanks to my supervisors Bernd Möbius and Jürgen Trouvain, for their guidance, patience, and unwavering support throughout the entire research process. Their expertise and valuable feedback have been instrumental in shaping the direction of this study. I feel very lucky to say that they made my time during this project very enjoyable and full of laughs. A big thank you to Joakim Gustafsson for agreeing to be my external reviewer. Thank you to Johannah O'Mahony and Éva Székely for being such wonderful collaboration partners.

I would also like to thank my fun-etics friends and colleagues in Saarbrücken. Ivan Yuen for your valuable insights. Omnia Ibrahim for endless kindness. And to my closest partners-in-crime: Beeke Muhlack and Raphael Werner. My time here will always be especially memorable because of you two. Beeke, thank you for always being so supportive and having my best interest in mind. You truly are the brightest witch of our age. Raphael, you have an uncanny ability to always make me laugh and smile. You have been like a brother.

I dedicate this work to my mother for her limitless love and inspiring me to continue education. Thank you to my dad for his constant support and pearls of wisdom. Lauren, thank you for always being available to chat and for cheering me up. Tim, you are the funniest dude and I am glad you made time to help me unwind. And last but definitely not least, a massive thank you to my wife Kitty. You have made everyday special and a joy. I love you.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project "Pause-internal phonetic particles in speech communication" (project number: 418659027; project IDs: MO 597/10-1 and TR 468/3-1).

As an associate member of SFB1102 "Information Density and Linguistic Encoding" and its integrated graduate program, I profited greatly from many opportunities for scientific training and exchange (project number: 232722074).

Abstract

Pause-internal phonetic particles (PINTs) comprise a variety of phenomena including: phonetic-acoustic silence, inhalation and exhalation breath noises, filler particles “uh” and “um” in English, tongue clicks, and many others. These particles are omni-present in spontaneous speech, however, they are under-researched in both natural speech and synthetic speech. The present work explores the influence of PINTs in small-context recall experiments, develops a bespoke speech synthesis system that incorporates the PINTs pattern of a single speaker, and evaluates the influence of PINTs on recall for larger material lengths, namely university lectures.

The benefit of PINTs on recall has been documented in natural speech in small-context laboratory settings, however, this area of research has been under-explored for synthetic speech. We devised two experiments to evaluate if PINTs have the same recall benefit for synthetic material that is found with natural material. In the first experiment, we evaluated the recollection of consecutive missing digits for a randomized 7-digit number. Results indicated that an inserted silence improved recall accuracy for digits immediately following. In the second experiment, we evaluated sentence recollection. Results indicated that sentences preceded by an inhalation breath noise were better recalled than those with no inhalation. Together, these results reveal that in single-sentence laboratory settings PINTs can improve recall for synthesized speech.

The speech synthesis systems used in the small-context recall experiments did not provide much freedom in terms of controlling PINT type or location. Therefore, we endeavoured to develop bespoke speech synthesis systems. Two neural text-to-speech (TTS) systems were created: one that used PINTs annotation labels in the training data, and another that did not include any PINTs labeling in the training material. The first system allowed fine-tuned control for inserting PINTs material into the rendered material. The second system produced PINTs probabilistally. To the best of our knowledge, these are the first TTS systems to render tongue clicks.

Equipped with greater control of synthesized PINTs, we returned to evaluating the recall benefit of PINTs. This time we evaluated the influence of PINTs on the recollection of key information in lectures, an ecologically valid task that focused on larger material lengths. Results indicated that key information that followed PINTs material was less likely to be recalled. We were unable to replicate the benefits of PINTs found in the small-context laboratory settings.

This body of work showcases that PINTs improve recall for TTS in small-context environments just like previous work had indicated for natural speech. Additionally, we’ve provided a technological contribution via a neural TTS system that exerts finer control over PINT type and placement. Lastly, we’ve shown the importance of using material rendered by speech synthesis systems in perceptual studies.

Ausführliche Zusammenfassung

Pauseninterne phonetische Partikeln (PINTs) umfassen eine Vielzahl von Phänomenen, wie z. B. akustisch-phonetische Stille, Ein- und Ausatmungsgeräusche, die Füllpartikeln "uh" und "um" im Englischen und "äh" und "ähm" im Deutschen sowie Zungenklicks. Diese Partikeln sind sowohl für natürliche als auch für synthetische Sprache weitgehend unerforscht. Die vorliegende Arbeit ist in drei Kapitel unterteilt.

Kapitel 3 befasst sich mit dem Einfluss von PINTs in synthetischer Sprache auf den Abruf von Ein-Satz-Laborstimuli. Kapitel 3 basiert auf den Veröffentlichungen Elmers et al. (2021a) und Elmers et al. (2021b). Frühere Arbeiten haben gezeigt, dass die Aufnahme von PINTs in natürliche Sprache Vorteile für die Wahrnehmung bringt. Daher ist es ein Hauptziel dieses Abschnitts, die Wahrnehmungsvorteile von PINTs in synthetischer Sprache zu bewerten. Es wurden zwei Experimente durchgeführt. Das erste Experiment verwendete ein konkatenatives Sprachsynthesesystem, um randomisierte siebenstellige Zahlen zu erzeugen. Das Erinnerungsvermögen der Teilnehmer wurde durch ein Wahrnehmungsexperiment bewertet, bei dem sie gebeten wurden, drei aufeinanderfolgende Ziffern in einem später präsentierten Strang zu ergänzen. Vor einigen der Ziffern wurden Pausen eingefügt, um deren Einfluss auf das Erinnerungsvermögen zu ermitteln. Das zweite Experiment, eine Replikationsstudie von Whalen et al. (1995), untersuchte das Erinnerungsvermögen der Teilnehmer an synthetische Sprache. In diesem Experiment wurde ein konkatenatives Synthesesystem zur Erzeugung von Sätzen verwendet. Vor einigen Sätzen wurden Einatmungsgeräusche eingefügt, um das Erinnerungsvermögen an die Satzinhalte zu bewerten. Insgesamt deuteten diese Experimente darauf hin, dass synthetisierte PINTs in Laborumgebungen mit nur einem Satz das Erinnerungsvermögen verbessern können. Die nächsten beiden Absätze beschreiben die beiden Experimente ausführlicher.

Evaluierung der Wirkung von Pausen auf das Erinnerungsvermögen an Zahlen bei synthetischer Sprache Diese Studie untersucht die Auswirkungen eingefügter Stille auf das Ziffernerinnerungsvermögen bei synthetisierter Sprache. Die Teilnehmer nahmen an einem Wahrnehmungsexperiment teil, bei dem sie eine siebenstellige Zufallszahl hörten, die von einem Sprachsynthesesystem wiedergegeben wurde. Bei einigen Stimuli wurde vor einer der Ziffern eine Pause (200 ms oder 500 ms lang) eingefügt, während andere keine Pause enthielten. Unmittelbar nach jedem Stimulus wurden die Teilnehmer gebeten, eine fehlende Folge von drei benachbarten Ziffern zu ergänzen. Die Ergebnisse zeigen, dass die Erinnerungsgenauigkeit unmittelbar nach einer Pause verbessert wird. Außerdem fanden wir einen signifikanten Effekt bei einer Pausendauer von 500 ms, aber nicht bei einer Pausendauer von 200 ms. Bei der Untersuchung der Reaktionszeit stellten wir fest, dass sich die Reaktionszeit der Teilnehmer erhöhte, wenn eine Pause vorhanden war. Insgesamt zeigen die Ergebnisse, dass Pausen in der synthetischen Sprache eine Rolle spielen. Diese Forschung kann im Zusammenhang mit der Untersuchung von Pausen und pausen-internen Partikeln (z. B. Atemgeräusche) in synthetisierter Sprache und deren Auswirkungen auf menschliche Zuhörer betrachtet werden.

Einatmen: Atemgeräusche verbessern das Erinnerungsvermögen bei synthetischer Sprache Diese Studie greift Whalen et al. (1995) auf, indem sie englischsprachige Teilnehmer in einem Wahrnehmungsexperiment auswertet, um festzustellen, ob ihr Erinnerungvermögen durch das Einfügen von Atemgeräuschen in Sätze, die von einem Sprachsynthesesystem erzeugt werden, beeinflusst wird. Whalen fand eine Verbesserung des Erinnerungsvermögen für Sätze, denen ein Atemgeräusch vorangestellt war, im Vergleich zu Sätzen ohne Atemgeräusch. Während Whalen et al. (1995) die englischen Sätze mit Hilfe der Formantensynthese wiedergaben, verwenden wir ein modernes konkatenatives Synthesesystem. In der vorliegenden Studie wurden Einatmungen von drei verschiedenen Längen verwendet: 0 ms (kein Atemgeräusch), 300 ms (kurzes Atemgeräusch) und 600 ms (langes Atemgeräusch). Unsere Ergebnisse stimmten mit denen von Whalen et al. (1995) für die 600 ms-Bedingung überein, aber nicht für die 300 ms-Bedingung, was darauf hindeutet, dass nicht alle Inhalationen das Erinnerungsvermögen verbessern. In der vorliegenden Studie wurde auch ein signifikanter Effekt für die Satzlänge festgestellt, was zeigt, dass kürzere Sätze eine höhere Erinnerungsgenauigkeit aufweisen als längere Sätze. Insgesamt deutet die vorliegende Studie darauf hin, dass Atemgeräusche für das Erinnerungsvermögen an synthetisierte Sprache wichtig sind und dass sich Forscher bei zukünftigen Studien auf längere und komplexere Arten von Sprache, wie Absätze oder Dialoge, konzentrieren sollten.

Kapitel 4 untersucht die Erkennung von PINTs und entwickelt eine maßgeschneiderte Sprachsynthese basierend auf dem PINTs-Muster eines einzelnen Sprechers. Kapitel 4 basiert auf den Veröffentlichungen Elmers (2022) und Elmers et al. (2023). Im ersten Experiment wurde die Klassifizierungsgenauigkeit von PINTs unter Verwendung verschiedener maschineller Lernarchitekturen bewertet. Die verschiedenen Architekturen für maschinelles Lernen schnitten ähnlich ab, wobei einige PINTs erfolgreich klassifiziert wurden, während andere nicht klassifiziert werden konnten. Dies veranlasste uns, ein internes Annotationsschema zu entwickeln, um ein Synthesesystem zu schaffen, das eine Vielzahl von PINTs erzeugen kann. Es wurden zwei Text-to-Speech-Systeme (TTS) entwickelt: eines, das PINTs mithilfe von konkreten Labels erzeugt, und ein zweites System, das keine Labels in den Trainingsdaten enthält und PINTs auf probabilistische Weise erzeugt. Das erste System bietet Kontrolle über die Platzierung und den PINT-Typ und ist unseres Wissens nach das erste System, das Zungenklicks erzeugt. Außerdem wurde eine Wahrnehmungsstudie mit Stimuli durchgeführt, die von dem gelabelten System erzeugt wurden. Insgesamt stellt der zweite Teil einen technologischen Beitrag dar und zeigt, dass Sprachsynthesesysteme, die natürliche Phänomene einbeziehen, leistungsstarke Werkzeuge für die Erstellung und Auswertung von manipuliertem Versuchsmaterial sein können. Die nächsten beiden Absätze beschreiben die beiden Experimente ausführlicher.

Vergleich von Erkennungsmethoden für pauseninterne Partikeln In dieser Studie wurden verschiedene Architekturen des maschinellen Lernens zur Klassifizierung von PINTs untersucht, wie z. B. Füllpartikeln (FPs), Atemgeräusche und Zungenklicks. Viele dieser PINTs treten gemeinsam auf, und durch die gleichzeitige Modellierung dieser PINTs soll die Klassifizierungsgenauigkeit auch für die umgebenden PINTs verbessert werden. Für die Modellierung wurde eine annotierte Teilmenge aus einem deutschen Spontansprachkorpus verwendet. Mel-Frequenz-Cepstrum-Koeffizienten wurden als Eingaben verwendet, um PINTs mit drei Arten von neuronalen Netzen zu modellieren: ein allgemeines neuronales Netz, ein konvolutionelles neuronales Netz und ein rekurrentes neuronales Netz. Die Modelle verwendeten die gleichen Hyperparameter, die gleiche Anzahl von Schichten und die gleiche Anzahl von Neuronen für diese Schichten, sodass der Schwerpunkt auf die Modellarchitektur gelegt wurde. Es wurde erwartet, dass das rekurrente neuronale Netz am besten abschneiden würde, da es in der Lage ist, zeitliche Informationen zu erfassen. Alle Modelle schnitten jedoch ähnlich ab. Die Modelle schnitten am besten bei der Klassifizierung stiller Segmente ab, gefolgt von Ein- und Ausatmungen. Allerdings gelang es allen Modellen nicht, FPs und Klicks genau zu klassifizieren, was darauf hindeutet, dass die gleichzeitige Modellierung von PINTs nicht immer die Genauigkeit für umgebende PINTs verbessert. Diese Ergebnisse deuten darauf hin, dass eine genaue Klassifizierung eher von der Quantität und Qualität der Annotation als von der Modellarchitektur abhängt. Die wichtigsten Beiträge dieser Arbeit sind die gleichzeitige Klassifizierung mehrerer PINTs und die Verbesserung der Klassifizierung von PINTs für die deutsche Sprache.

Synthese nach ein paar PINTs: Untersuchung der Rolle von pauseninternen phonetischen Partikeln in der Sprachsynthese und -wahrnehmung Pauseninterne phonetische Partikel, wie z. B. Stille, Atemgeräusche, Füllpartikeln, Zungenschnalzen und Zögern sind in der natürlichen Sprache weit verbreitet. Diese Partikeln spielen eine wichtige Rolle in der Sprachwahrnehmung, werden aber in der Sprachsynthese selten modelliert. Wir haben zwei TTS-Systeme entwickelt: 1) ControlledPINT, ein Modell, das PINT-Labels in die Trainingsdaten einbezog, und 2) AutoPINT, ein Modell, das keine PINT-Labels in die Trainingsdaten einbezog. Beide Modelle produzierten weniger PINTs und hatten eine geringere Gesamtdauer der PINTs als die natürliche Sprache. Das gelabelte Modell erzeugte mehr PINTs und hatte eine längere Gesamtdauer der PINTs als das Modell ohne Labels. In einem Hörexperiment mit dem gelabelten Modell haben wir den Einfluss verschiedener PINT-Kombinationen auf die Wahrnehmung der Sprechersicherheit untersucht. Wir testeten vier Bedingungen, die durch das ControlledPINT-Modell generiert wurden: eine "flüssige" Bedingung ohne PINTs-Material, eine Bedingung mit langem Schweigen, eine Bedingung mit Füllpartikeln und eine kombinierte Bedingung, die Schweigen, Füllpartikel "um", Zungenklick und Einatmung beinhaltete. Die Bedingung ohne PINTs wurde als signifikant selbstsicherer wahrgenommen als die PINTs-Bedingungen, was darauf hindeutet, dass wir durch die Einbeziehung von PINTs verändern können, wie selbstsicher TTS-Sprecher wahrgenommen werden. Die drei Bedingungen mit PINTs schnitten insgesamt ähnlich ab, wobei die Bedingung mit langem Schweigen geringfügig besser abschnitt als die Bedingung mit Füllpartikeln, die wiederum geringfügig besser abschnitt als die Kombinationsbedingung. Diese Ergebnisse zeigen, dass das Einfügen von PINTs in synthetische Sprache dazu verwendet werden kann, den Klang des Materials zu beeinflussen. Darüber hinaus unterstreicht diese Studie, dass der Output von TTS-Systemen für die Untersuchung von Forschungsfragen im Bereich der Sprachwissenschaft genutzt werden kann. Da TTS-Anwendungen, wie z. B. Konversationssysteme, zunehmend in der Lage sind, eine lebensnahe Kommunikation zu ermöglichen, muss die Rolle dieser spontanen Sprachphänomene besser verstanden werden und Teil der generativen Modellierung werden.

Kapitel 5 vergleicht die Verwendung von PINTs zwischen Universitätsdozenten und englischsprachigen Testmaterialien und verbindet den perzeptuellen Erinnerungsfokus aus Kapitel 3 mit dem maßgeschneidert Sprachsynthesesystem aus Kapitel 4. Kapitel 5 basiert auf den Publikationen Elmers & Trouvain (2022), Elmers (2023) und Elmers & Székely (2023). Zunächst verglichen wir die Verwendung von PINTs aus Vorlesungen der Yale University mit dem TOEFL iBT Hörverstehensübungsabschnitt. Wir stellten fest, dass die PINTs fast 1/3 der Gesamtzeit in den Yale-Vorlesungen ausmachten. Dieses Ergebnis deutet darauf hin, dass untersucht werden muss, wie PINTs den Abruf von Schlüsselinformationen in Vorlesungen beeinflussen. Anstatt sich auf Einzelsatz-Laborstimuli zu konzentrieren, konzentrierte sich das Kapitel 5 auf die Evaluierung der Erinnerungseffekte von PINTs in realen Bildungssituationen. Als nächstes führten wir ein Wahrnehmungsexperiment durch, bei dem natürliche Sprache aus Universitätsvorlesungen verwendet wurde. Die Teilnehmer waren sowohl englische Muttersprachler als auch Nicht-Muttersprachler. Die Teilnehmer hörten dreiminütige Abschnitte, die aus Vorlesungen in voller Länge extrahiert wurden, und beantworteten Multiple-Choice-Fragen. Einigen Informationen, die für die Beantwortung der Fragen entscheidend waren, wurden PINTs vorangestellt. Die Ergebnisse zeigten, dass Inhalte, denen PINT-Material unmittelbar vorausging, mit geringerer Wahrscheinlichkeit abgerufen wurden. Im dritten Experiment wurde das gleiche Versuchsparadigma wie im zweiten Experiment verwendet, allerdings mit synthetischer Sprache. Die Audioinhalte wurden genauso wiedergegeben wie die natürliche Sprache. Auch hier zeigte sich, dass Informationen, denen PINTs unmittelbar vorausgingen, weniger wahrscheinlich abgerufen wurden. Insgesamt waren diese Experimente nicht in der Lage, die Vorteile von PINTs, die in Einzelsatz-Laborsituationen gefunden wurden, in realen Vorlesungsszenarien zu replizieren. Wichtig ist, dass in diesen Studien nur das PINT-Profil eines einzelnen Sprechers untersucht wurde. Die Muttersprache der Teilnehmer hatte in keinem der beiden Experimente einen signifikanten Einfluss auf das Erinnerungsvermögen. Dies ist ein positives Ergebnis, das darauf hindeutet, dass die Einbeziehung von PINTs-Material in ein Sprachsynthesesystem keine anderen Auswirkungen auf nicht-muttersprachliche Hörer hat als auf muttersprachliche Hörer.

Diese Arbeit dient der Erforschung der komplexen Effekte von PINTs. Zukünftige Arbeiten sollten weiterhin eine Vielzahl von PINT-Sprecherprofilen evaluieren und untersuchen, wie die Verwendung von PINTs die Leistung in Bildungsumgebungen beeinflusst, insbesondere in Universitätsumgebungen, in denen Studenten mit unterschiedlichem Sprachhintergrund auf der Grundlage ihrer Erinnerung an Vorlesungsmaterial bewertet werden. Die nächsten drei Absätze beschreiben die Experimente ausführlicher.

Vergleich der PINTs in Universitätsvorlesungen In dieser Studie wurde die Verwendung von PINTs in fünf Universitätsvorlesungen aus Open Yale Courses (2007b) mit dem TOEFL iBT Hörtest für englischsprachige Vorlesungen verglichen. Die folgenden PINTs wurden annotiert: Schweigen, Einatmen, Ausatmen, "uh", "um", Zungenklicks und eine Kategorie "andere". Insgesamt wurden für die Yale-Vorlesungen Material von 5 Stunden (1 Stunde pro Sprecher) und für die TOEFL-Vorlesungen 15 Minuten annotiert. Die Yale-Vorlesungen wurden während eines dreimonatigen Semesters aufgezeichnet, so dass die Annotationen zu Beginn, in der Mitte und am Ende des Semesters gemacht wurden, um die Variation der PINTs zwischen den Dozenten zu vergleichen. Die Ergebnisse zeigten, dass die PINTs bei den Yale-Vorlesungen 30% der Gesamtzeit ausmachten und bei den TOEFL-Vorlesungen 20%. Bei den Yale-Vorlesungen wurde bei den verschiedenen Sprechern eine unterschiedliche Verwendung von PINTs in Bezug auf die Art der PINTs, die Anzahl, die Dauer und die Häufigkeit festgestellt. Allerdings waren die Dozenten während des gesamten Semesters in ihrer PINT-Verwendung konsistent, was auf minimale Variationen innerhalb der Dozenten hindeutet. Der hohe Anteil an PINTs in Vorlesungen deutet darauf hin, dass weitere Arbeiten durchgeführt werden sollten, um zu untersuchen, wie diese Partikeln den Abruf von Vorlesungsmaterial beeinflussen.

Einfluss von Pausenpartikeln auf das Erinnern von Vorträgen Diese Studie untersucht den Einfluss von PINTs auf die Erinnerung an natürliche Sprache bei muttersprachlichen und nicht-muttersprachlichen Hörern des Englischen. Die Teilnehmer waren 45 monolinguale englische und 45 deutsche L1-Hörer, die Abschnitte aus Universitätsvorlesungen in englischer Sprache hörten und inhaltliche Fragen beantworteten. Es wurden drei Versionen der Vorlesungsstimuli erstellt: eine unmanipulierte Originalversion, eine "Silence"-Version und eine "No PINTs"-Version, bei der alle PINTs einschließlich der Stille entfernt wurden. In der Original- und der "Silence"-Version wurde die Hälfte der Schlüsselinformationen durch PINT-Material eingeleitet. Die Ergebnisse zeigten, dass das Material, dem PINTs vorangestellt waren, mit geringerer Wahrscheinlichkeit abgerufen wurde. Außerdem war die Erstsprache der Teilnehmer für das Verstehen des Sprechers nicht von Bedeutung. Allerdings schnitten englische Hörer in der Bedingung "keine PINTs" tendenziell besser ab, während deutsche Hörer in der Originalbedingung tendenziell besser abschnitten. In dieser Studie konnte der in den Laborexperimenten mit Einzelsätzen gefundene Erinnerungsvorteil von PINTs nicht repliziert werden. Die Interaktion zwischen PINTs und dem Abruf von Informationen ist komplex, insbesondere in realen Vorlesungsszenarien. Diese Arbeit zielt darauf ab, das Verständnis von PINTs zu verbessern und zu zeigen, wie sie sowohl muttersprachliche als auch nicht-muttersprachliche Hörer in Bildungssituationen beeinflussen.

Die Auswirkung von pauseninternen phonetischen Partikeln auf den Abruf in synthetisierten Vorlesungen Wir untersuchten die Auswirkung von PINTs auf das Erinnerungsvermögen von englischen Muttersprachlern und Nicht-Muttersprachlern in einem Hörexperiment mit synthetisiertem Material, das eine Universitätsvorlesung simulierte. Mit Hilfe eines neuronalen Sprachsynthesizers, der auf aufgezeichnete Vorlesungen mit PINTs-Kommentaren trainiert wurde, erzeugten wir drei verschiedene Bedingungen: eine Basisversion, eine "Silence"-Version, bei der nicht-stille PINTs durch Stille ersetzt wurden, und eine "No PINTs"-Version, bei der alle PINTs, einschließlich Stille, entfernt wurden. Die Hälfte der Teilnehmer wurde darüber informiert, dass sie computergenerierte Audiodaten hörten, während die andere Hälfte darüber informiert wurde, dass die Audiodaten mit einem Mikrofon von schlechter Qualität aufgenommen worden waren. Zusätzlich haben wir die Meinungen der Teilnehmer zu den Audios, wie z. B. ihr Interesse, mit einem Fragebogen erhoben. Es zeigte sich, dass weder die Bedingung noch die Muttersprache der Teilnehmer einen signifikanten Einfluss auf das Gesamtergebnis hatten, und dass das Vorhandensein von PINTs vor kritischen Informationen einen negativen Effekt auf die Erinnerung hatte. Das Interesse der Teilnehmer an den Audioinhalten wirkte sich signifikant positiv auf die Erinnerungsleistung aus. Diese Studie unterstreicht die Bedeutung der Berücksichtigung von PINTs für Bildungszwecke in Sprachsynthesesystemen.

Publications

Parts of this dissertation have appeared in the following publications:

Elmers, M. (2022). Comparing detection methods for pause-internal particles. In *Proc. 33rd Conference Elektronische Sprachsignalverarbeitung (ESSV '22)* (pp. 204–211). URL: https://www.essv.de/paper.php?id=1160.

Elmers, M. (2023). Pause particles influencing recollection in lectures. In *Proc. 20th International Congress of Phonetic Sciences (ICPhS '23)* (pp. 37–41). Prague. URL: https://guarant.cz/icphs2023/85.pdf.

Elmers, M., O'Mahony, J., & Székely, É. (2023). Synthesis after a couple PINTs: Investigating the role of pause-internal phonetic particles in speech synthesis and perception. In *Proc. Interspeech 2023* (pp. 4843–4847). URL: https://doi.org/10.21437/Interspeech.2023-2178.

Elmers, M., & Székely, É. (2023). The impact of pause-internal phonetic particles on recall in synthesized lectures. In *Proc. 12th ISCA Speech Synthesis Workshop (SSW 12)* (pp. 204–210). URL: https://doi.org/10.21437/SSW.2023-32.

Elmers, M., & Trouvain, J. (2022). Pause-internal particles in university lectures. Poster presentation at 18th Phonetik & Phonologie (P&P '22). URL: https://mikeyelmers.github.io/publications/elmers_pp22_poster.pdf.

Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. (2021a). Evaluating the effect of pauses on number recollection in synthesized speech. In *Proc. 32nd Conference Elektronische Sprachsignalverarbeitung (ESSV '21)* (pp. 289–295). URL: https://www.essv.de/paper.php?id=1131.

Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. (2021b). Take a breath: Respiratory sounds improve recollection in synthetic speech. In *Proc. Interspeech 2021* (pp. 3196–3200). URL: https://doi.org/10.21437/Interspeech.2021-1496.

Contents

Chapter 1

Introduction

1.1 Motivation

Pause-internal phonetic particles (PINTs) comprise a variety of phenomena including: phonetic-acoustic silence, inhalation and exhalation breath noises, filler particles “uh” and “um” in English, tongue clicks, and many others. The benefit of PINTs on recall has been documented in natural speech in small-context laboratory settings, however, this area of research has been under-explored for synthetic speech. Therefore, a major goal of this thesis was to evaluate the perceptual benefits of PINTs on recall in synthetic speech.

Speech synthesis systems often do not provide much freedom in terms of controlling PINT type or location. Therefore, we endeavoured to develop bespoke speech synthesis systems that produce a variety of PINTs. Two neural text-to-speech (TTS) systems were created: one that used PINTs annotation labels in the training data, and another that did not include any PINTs labeling in the training material. The first system allowed fine-tuned control for inserting PINTs material into the rendered material. The second system produced PINTs probabilistally. To the best of our knowledge, these are the first TTS systems to render tongue clicks. Importantly, these systems showcased that the output of TTS systems can be used to investigate research questions in the speech science field. As TTS applications, such as conversational systems, become increasingly capable of facilitating lifelike communication, the roles of these spontaneous speech phenomena will need to be better understood and become part of the generative modeling.

Using one of our developed TTS systems, we returned to evaluating the recall benefit of PINTs. This time we evaluated the influence of PINTs on the recollection of key information in lectures, an ecologically valid task that focused on larger material lengths. This work improved the understanding of PINTs and how they influence listeners in educational settings. Furthermore, this work highlighted the importance of considering PINTs for educational purposes in speech synthesis systems.

This body of work showcases that PINTs improve recall for TTS in small-context environments just like previous work had indicated for natural speech. Additionally, we've provided a technological contribution via a neural TTS system that exerts finer control over PINT type and placement. We've shown the importance of using material rendered by speech synthesis systems in perceptual studies. Ultimately, the motivation behind this body of work is to improve the generation of synthetic PINTs and evaluate the effects of PINTs in laboratory and educational settings.

1.2 Research Questions

This thesis explores the following research questions:

1. What are the perceptual effects of PINTs in synthesized speech?
    - How do PINTs influence recall in single-sentence laboratory experiments?
    - How do PINTs influence recall in larger material lengths (i.e., university lectures)?
        - Do PINTs influence native and non-native listeners differently?
2. How well do different detection methods classify PINTs?
3. How much control can researchers have over the insertion of PINTs for speech synthesis systems?
4. How often do PINTs occur in lecture environments?

1.3 Structure

This body of work is divided into three main chapters. Chapter 3 focuses on the influence of PINTs in synthetic speech on the recall of single-sentence laboratory stimuli. Chapter 3 is based on the publications Elmers et al. (2021a) and Elmers et al. (2021b). The first experiment used a concatenative speech synthesis system to generate randomized 7-digit numbers. Participant's recollection was evaluated via a perceptual experiment where they were asked to recall three consecutive missing digits. Silences were inserted before some of the digits to evaluate their influence on recollection. The second experiment partially replicated Whalen et al. (1995) by evaluating participant's recollection of synthetic audio. This experiment used a concatenative synthesis system to generate sentences. Inhalation breath noises were inserted before some of the sentences to evaluate recollection of the sentence contents. Overall, these experiments indicated that in single-sentence laboratory environments synthesized PINTs can improve recall.

Chapter 4 investigated the detection of PINTs and developed bespoke speech synthesis systems based on the PINTs pattern of a single speaker. Chapter 4 is based on the publications Elmers (2022) and Elmers et al. (2023). The first experiment evaluated the classification accuracy of PINTs using different machine learning architectures. The different machine learning architectures performed similarly, successfully classifying some PINTs, while failing to classify others. This led us to develop an internal annotation schema with the purpose of creating a synthesis system capable of generating a variety of PINTs. Two TTS systems were developed: one that produced PINTs with overt labels, and a second system that did not include labels in the training data, which produced PINTs probabilistically. Additionally, a perception study was conducted using stimuli generated from the labeled system. Overall, the second part provides a technological contribution and showcases that speech synthesis systems, that include natural phenomena, can be powerful tools for creating and evaluating manipulated experimental material.

Chapter 5 compares the PINTs usage between university lecturers and English-language test materials. Chapter 5 also merges the perceptual recall focus from chapter 3 with the custom speech synthesis system from chapter 4. Chapter 5 is based on the publications Elmers & Trouvain (2022), Elmers (2023), and Elmers & Székely (2023). First, we compared the PINTs usage from Yale University lectures to the TOEFL iBT lecture listening practice section. We found that PINTs comprised almost 1/3 of the total time in the Yale lectures. This finding indicated a need for evaluating how PINTs influence the recall of key lecture information. Rather than focusing on single-sentence laboratory stimuli, chapter 5 focuses on evaluating the recall effects of PINTs in real-world educational settings. Next, we conducted a perceptual experiment that used natural speech from university lectures. Participants included both native and non-native speakers of English. Participants heard three-minute sections, that were extracted from full length length lectures, and then answered multiple-choice questions. Some of the information that was critical to answering the questions was preceded by PINTs. Results revealed that content immediately preceded by PINTs material was less likely to be recalled. The third experiment used the same experimental paradigm from experiment two, except with synthesized speech. The audio contents were rendered the same as the natural speech. Again, we found that information immediately preceded by PINTs was less likely to recalled. Overall, these experiments were not able to replicate the benefits of PINTs found in single-sentence laboratory settings in real-world lecture scenarios. The participants' native language did not significantly influence their recall in either experiment.

Chapter 2

Background

Pauses exhibit a great variety in their type and focus. For example, Trouvain & Werner (2022) describe four kinds of pauses: 1) articulation pauses, 2) listener pauses, 3) transition pauses during conversational turn-taking, and 4) pauses during connected speech. This thesis focuses on the fourth definition of pauses and investigates the pauses found within monologic speech. Pause-internal phonetic particles (PINTs) describe a variety of phenomena, such as acoustic-phonetic silence, breath noises (i.e., inhalations and exhalations), filler particles (FPs) like "uh" or "um", and tongue clicks. PINTs, in general, have a wide variety of functions and applications. PINTs are sometimes referred to as non-verbal vocalizations (NVVs), hesitation phenomena, or disfluencies. However, each of these classifications has its own definition and focus, often without a consensus amongst researchers. Therefore, this work will exclusively use the term PINTs, with definitions and focus provided in this chapter. Specifically, we investigated the influences of PINTs in recall, their inclusion in text-to-speech (TTS) synthesis systems, and their general educational applications. This chapter contains a section dedicated to each of the PINTs investigated in this work.

2.1 Silences

Silence segments (henceforth silences) refer to periods of acoustic-phonetic silence that are silent in production, but not in transmission. We use a definition of silences similar to Belz & Trouvain (2019). In other words, silences are phases absent of other phonetic particles such as breath noises, clicks, laughter, etc. Figure 2.1 provides an annotated silence example with spectrogram and waveform information. Even when the spectrogram and waveform both show an absence of sound, silences are not always obvious to listeners. MacIntyre & Scott (2022) found that participants could not reliably detect gaps of silence until the silence exceeded 440 ms, and that silence gaps below 200 ms were only correctly detected at chance levels. Goldman-Eisler

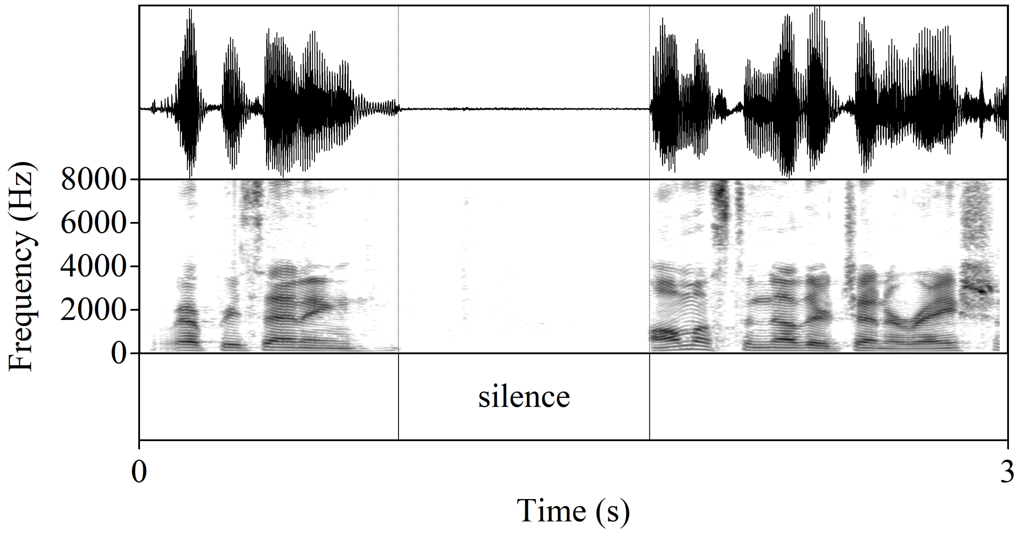


Figure 2.1: Example waveform and spectrogram for silence annotation.

(1961) found that a majority of silences in discussions are less than 1 second, and that very few silences are longer than 2 seconds. Silence durations between sentences for Japanese newscasters have been modeled using the preceding and following prosodic information (Nakamura et al., 2020). Additionally, both intonation phrase length and prosodic structure have been shown to affect the duration of silences (Krivokapić, 2007). For example, silences have an important role in breaking up speech. Goldman-Eisler (1958) found that silences are often found before key information words, and after redundant information words.

Silence, like many of the other PINTs investigated in this work, are highly individualistic with a mixture of influences. Silence rates also showcase unique values based on the gender, ethnicity, and geographic region of the speaker (Kendall, 2009). Similarly, silences exhibit a variety of lengths due to speaker idiosyncrasies, silence types, and dialogues (Fors, 2015). Silences are also influenced via diachronic change and style differences. For example, Trouvain (2011) showed that the extensive use of silences found in 1970s television commentaries is no longer common in modern times, and differs from the silence usage of radio commentators. When comparing political interviews, casual interviews, and political speeches, Duez (1982) also found stylistic differences with the total silence duration being 50% longer in political speeches than for the interviews.

2.2 Breath Noises

This work focuses on breath noises that occur during speech, such as inhalations and exhalations. Breath noises, like silences, also vary by speaker. Breath noises, for

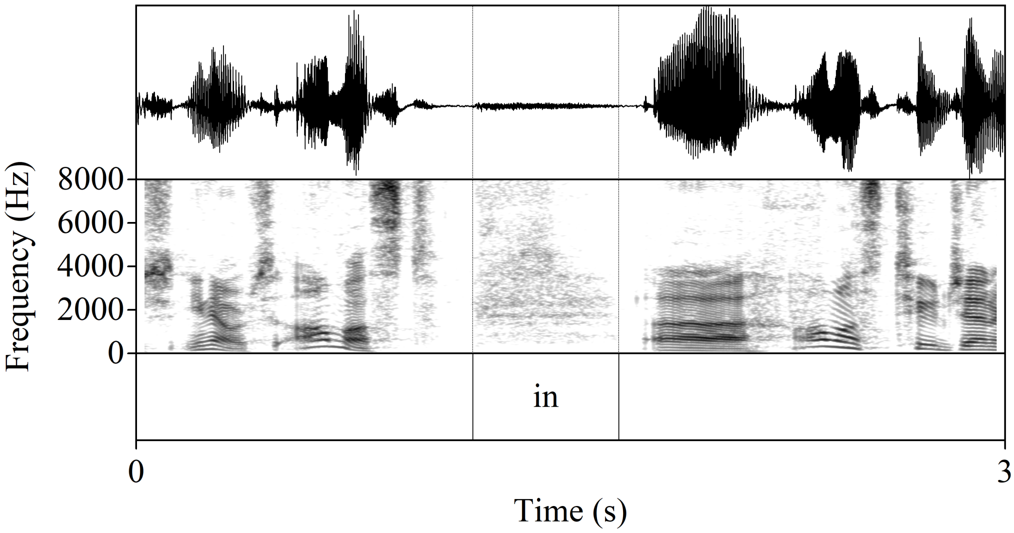


Figure 2.2: Example waveform and spectrogram for inhalation breath noise annotation.

example, can mark speaker individuality (Kienast & Glitza, 2003) or indicate formality in Korean (Winter & Grawunder, 2012). Breath noises are often associated with syntactic-prosodic breaks (Trouvain et al., 2020). Trouvain et al. (2020) also shows that many pauses contain inhalation noises, that inhalations are often low in intensity, and vary greatly in duration. Figure 2.2 and Figure 2.3 provide an annotated example for an inhalation and exhalation, respectively. Inhalations, in particular, can provide perceptual cues to listeners. While evaluating single sentences, Whalen & Kinsella-Shaw (1997) found a positive correlation between the duration of an inhalation and the length of the upcoming sentence. Fuchs et al. (2013) similarly found that the intensity and duration of breath noises are influenced by speech planning, and can help the listener predict the amount of upcoming material. Breath noises also frequently display a relationship with prosodic breaks during turn-taking. Overall, speakers tend to economize breathing in their speech to their communicative needs (Włodarczak et al., 2015).

2.3 Filler Particles

Filler particles (FPs) are another common PINT type, often with different realizations depending on the language. Examples of FPs include: “uh” and “um” in English, “äh” and “äm” in German, and “eto” and “ano” in Japanese. Figure 2.4 and Figure 2.5 show annotated examples for the English FPs “uh” and “um”, respectively. FPs have a number of functions. For example, FPs exhibit communicative functions for turn-taking and maintaining the floor (Clark & Fox Tree, 2002), and as a sociolinguistic

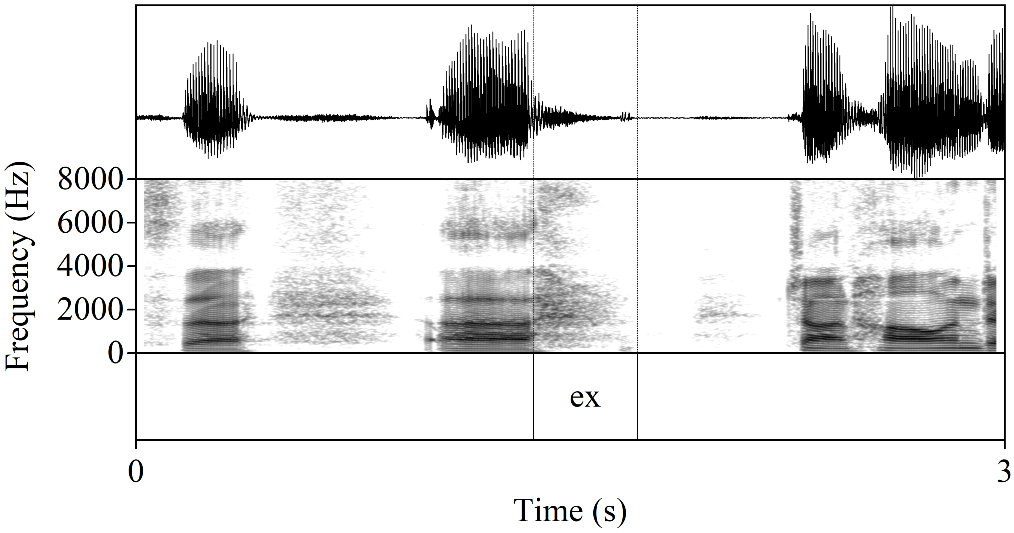


Figure 2.3: Example waveform and spectrogram for exhalation breath noise annotation.

identifier (Fruehwald, 2016). FPs can function as markers for forensic purposes (Braun & Rosin, 2015; Muhlack et al., 2023), and have technological applications for forensic voice comparison (Hughes et al., 2016). Laserna et al. (2014) showed that FP usage correlated with age but not gender. FPs usage also varies between languages. For example, Di Napoli (2020) found that FPs in Italian were far less common than prolongations, which differs to what Eklund (2001) found for Swedish. Muhlack (2023) found that English speakers preferred vocalic-nasal FPs (um), while Spanish speakers used vocalic FPs (uh). de Leeuw (2007) found language-related differences in the realization, location, and frequency of FPs in English, German, and Dutch. In a longitudinal study, de Boer et al. (2022) evaluated L1 (Dutch) and L2 (English) FPs at two time points (2.5 years), and found that FP spectral characteristics were consistent. While there are high levels of variability for FPs between speakers and languages, Lo (2020) showed that FP usage and acquisition for simultaneous bilingual speakers is also influenced by their linguistic surroundings. Using a map task where participants would play as both an instructor and instructee, Belz & Klapi (2013) found the role was a significant factor for determining L1 and L2 FP usage. Specifically, Belz & Klapi (2013) found that L1 and L2 speakers have different FP behaviors, with L2 speech incorporating longer silences after FPs, compared to the shorter silences after FPs in L1. Silber-Varod et al. (2020) found that with task-oriented dialogues, speakers differed significantly in their FP usage, however, there was no difference when the speaker switched roles.

While evaluating the presence of FPs at Japanese sentence and clause boundaries, Watanabe et al. (2006) found that the ratio of FPs was dependent on the complexity

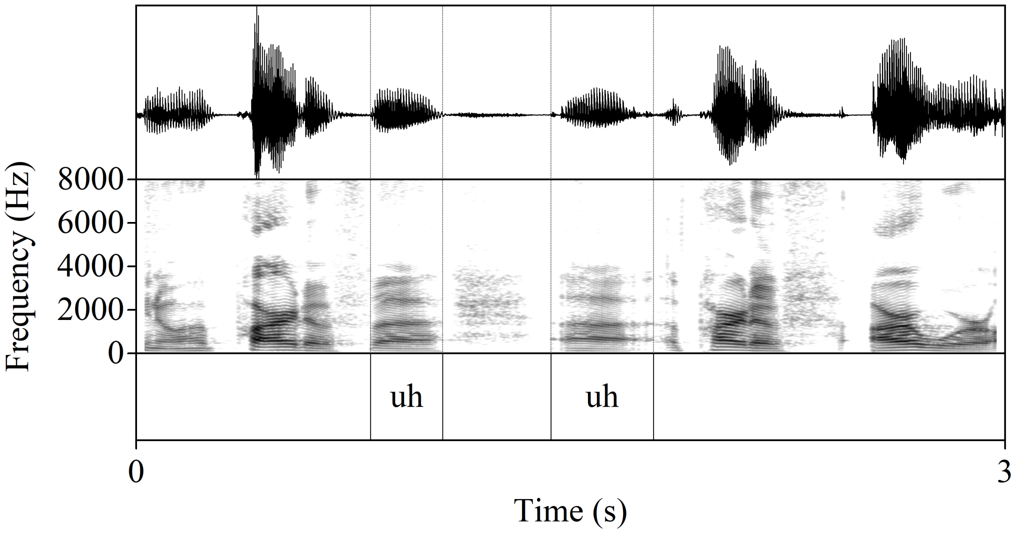


Figure 2.4: Example waveform and spectrogram for filler particle "uh" annotation.

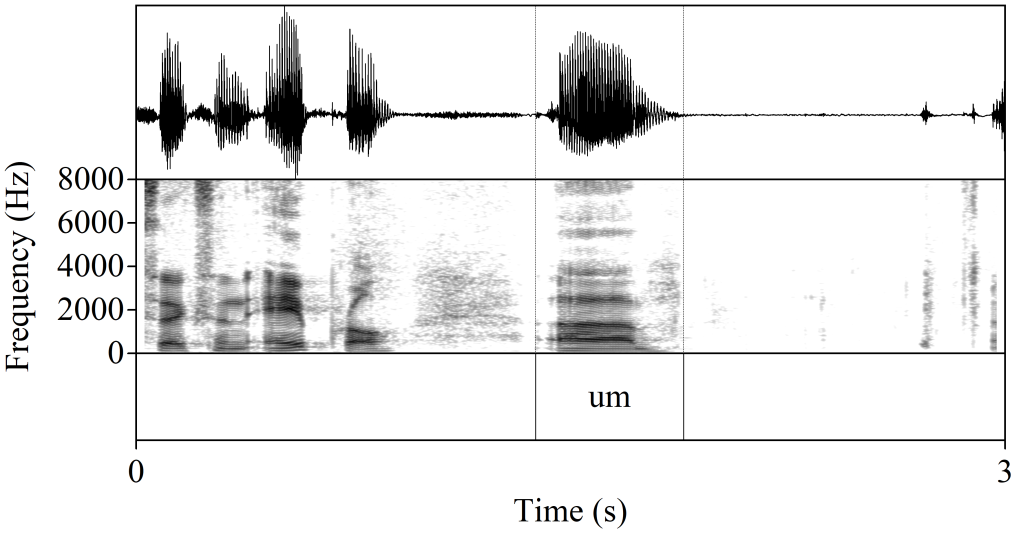


Figure 2.5: Example waveform and spectrogram for filler particle "um" annotation.

of the upcoming material, with higher ratios found before complex clause boundaries. Barr & Seyfeddinipur (2010) found that listeners expected new information when the speaker used “um”, but the expectation was dependent on what was new information for both the listener and speaker. Watanabe et al. (2008) found that Japanese listeners had a faster response time when phrases were led by a FP, compared to when no FP or silence was present. Conversely, Kosmala & Morgenstern (2019) did not find a relationship between the rate of FPs and perceived question difficulty. Thus asserting FPs occurred usually in an initial position, and are used mostly for planning and for buying time.

In terms of public-speaking, FPs are often criticized, and orators recommend trying to remove FPs from presentations. However, listener awareness of FP usage isn’t always accurate and can be influenced. For example, Niebuhr & Fischer (2019) found that listeners are less aware of both short and nasalized FPs, and that when speaker’s used these types of FPs their oratory skills were rated higher. Transcribers also make mistakes and often fail to transcribe FPs and other spontaneous speech phenomena (Zayats et al., 2019). Overall, FP usage incorporates personal characteristics, and FPs can be used for the prediction of upcoming information.

2.4 Tongue Clicks

Tongue clicks (henceforth clicks) involve velar ingressive suction with the back of the tongue near the velum and the blade of the tongue near the front of the mouth, potentially between the teeth or at the alveolar ridge (Ogden, 2020). Figure 2.6 shows an example of a click. Similar to other PINTs, clicks can also serve a variety of functions. One unique function of clicks is the their ability to be indirect. For example, clicks can refer to socially inappropriate topics, such as swearing, self-aggrandizing, or sexual innuendos (Ogden, 2020). Additionally, it is possible to combine clicks with gestures for additional effects or new meanings (e.g., click + wink). Clicks can also have other social functions, such as indicating assertive or authoritative positions in Irish English (Schulte, 2020). Clicks can function as discourse markers that index a new sequence or signal formulation difficulties (Trouvain & Malisz, 2016). Pinto & Vigil (2019) found that search clicks appear more frequently near nouns or noun phrases, and that clicks might function as a hedge in a search event. However, Vigil & Pinto (2020) found that listeners were unable to detect click sounds, and that it might not be ideal to have long periods of silence. Since clicks can influence the perception of speech, making it seem less dynamic or more more monotone.

2.5 Combination

So far, each particle has been shown in isolation, however, real-word examples are usually more complex and involve the co-occurrence of multiple PINTs (see Figure 2.7).

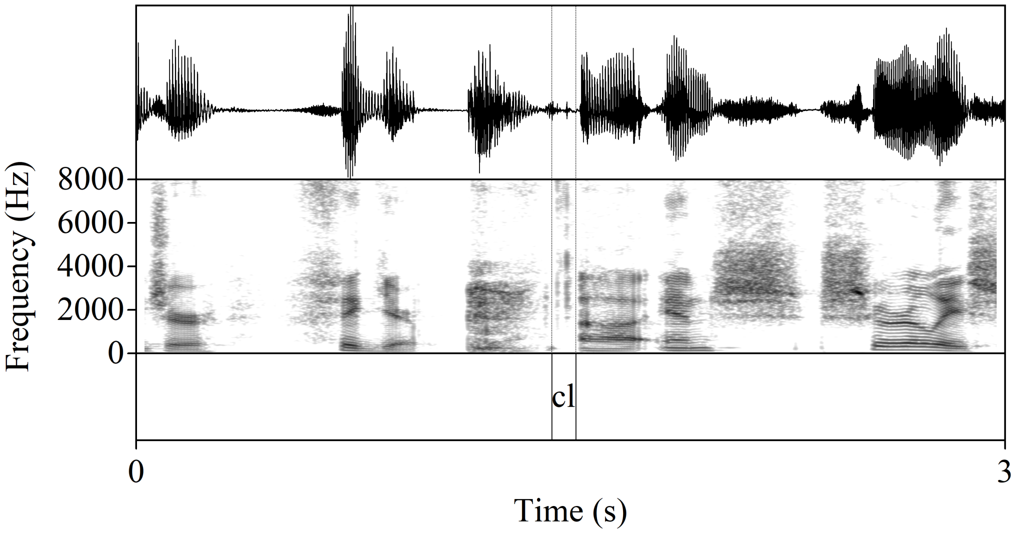


Figure 2.6: Example waveform and spectrogram for tongue click annotation.

For example, Zellers (2022) found that FPs tended to appear next to search clicks while inhalations appeared next to turn-continuation effects. Ogden (2013) also found that clicks co-occurred in word-search, alongside other PINTs such as FPs, inhalations, and silences. Clicks that index new sequences occurred at different syntactic locations than clicks which index word searching. Moreno (2019) similarly found clicks were used in word search and co-occur with filler particles.

Silences often are found surrounding other PINTs. Silence duration is influenced by the type of FP, with silence durations being consistently shorter before FPs (Betz & Kosmala, 2019). Many short silences, known as edge silences, are found adjacent to other PINTs. Again, we also find that the distribution of co-occurring PINTs depends on the communicative purposes (i.e., style). For example, oral presentations include more pre-utterance clicks and inhalations (Kosmala, 2020). Adell et al. (2012) found that 60% of FPs were preceded by a silence, and that only 24% of FPs were followed by a silence. Speakers are able to exert their unique linguistic flair via their PINTs usage, resulting in variation both between and within speakers for PINTs. For example, some speakers heavily use FPs while others prefer not to use them, and some speakers use short or long silences during their planning, or incorporate other hesitations (Betz & Gambino, 2016). The use of both silences and FPs has been associated with honest speech, rather than deceptive speech (Benus et al., 2006).

2.6 Other PINTs

There are many other phenomena found within pauses that are not the focus of this work. The following PINTs were not evaluated: laughter, swallowing, coughing,

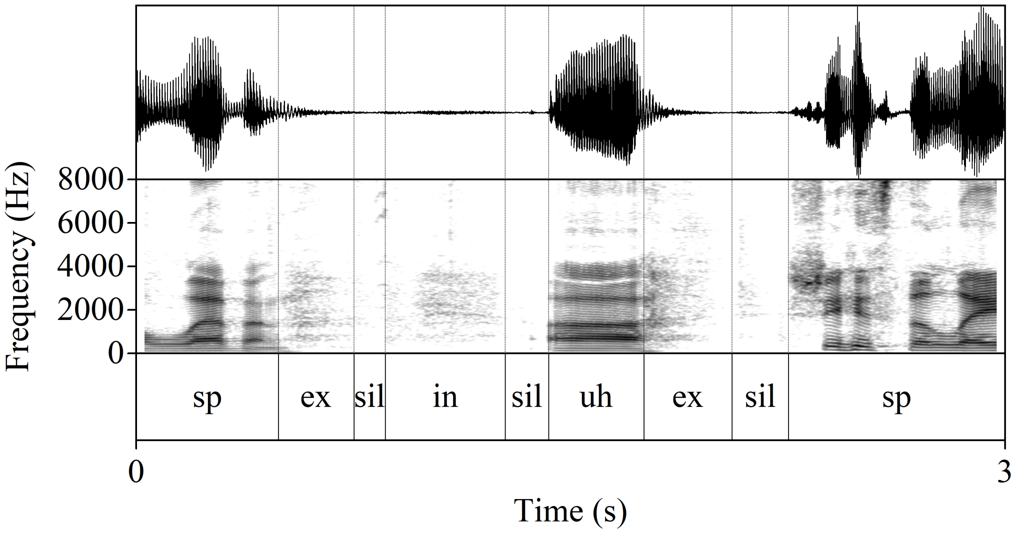


Figure 2.7: Example waveform and spectrogram for co-occuring PINTs annotations. Annotations include silence (sil), inhalation noise (in), exhalation noise (ex), filler particles (uh), and speech (sp).

throat clearing, sniffling, lip smacking, yawning, interjections (e.g., “wow” in English), back-channels (e.g., “uh-huh” in English), lexical fillers (e.g., “you know” in English), and discourse markers (e.g., “like” in English). There are other phenomena that are often associated with PINTs that also aren’t covered in this work such as: repairs (e.g. replacement of immediately preceding erroneous speech), repeats (e.g., the repetition of a single or series of words), false starts (e.g., start speech but quit before completion), and lengthenings (e.g., prolongation of part or parts of a word). A popular model for describing hesitation behavior was put forth by Levelt for self-monitoring and error repair (1983). This model explains how errors are detected, the incorporation of an editing term (e.g., silence or FP), and the repair. These lists are not exhaustive, especially as researchers become more granular in their annotations of these phenomena. Importantly, all of these phenomena have a richness and wealth to their exploration but are out of scope for this work.

2.7 PINTs in TTS

Speech synthesis systems display large amounts of variation in how they handle PINTs. For example, PINTs are often handled haphazardly, applying rudimentary punctuation-based heuristics for determining their location, frequency, and duration. Most modern TTS systems do not implement PINTs with appropriate placement and duration (Trouvain & Möbius, 2018), and fail to include any breath noises whatsoever. However, there are a number of notable exceptions (Braunschweiler & Chen, 2013; Székely et al., 2020).

PINTs are an important tool for helping listeners subdivide synthesized speech into easily understood pieces. When trying to model silences, Yang et al. (2014) found syntax, discourse, topic, and length all influenced the silence duration, showcasing some of the numerous factors that influences silence durations. Moreover, using a regression model, Yang et al. (2014) found that approximately 80% of silence duration variance could be explained using only syntax, discourse hierarchy, and post-boundary length, indicating that some factors contribute more than others. Similarly, Rose (2017) found that silences have a stronger association with syntatic and discourse planning than FPs. Since there is structure to the placement of FPs, they can be modelled for TTS systems (Dall et al., 2014a). Gustafson et al. (2021) was able to insert fillers found in spontaneous speech without changing the personality. Modeling the placement of silences is important for TTS, since accurate prediction can improve both naturalness and intelligibility (Braunschweiler & Maia, 2016). Wang et al. (2010) found that grammatically inappropriate breath noise durations and locations were more present in spontaneous speech tasks than passage reading in natural speech. However, this finding also has implications for modeling duration and location of breath noises in speech synthesis, and shows that simple heuristics are inadequate for modeling how breath noises occur in natural spontaneous speech. Moreover, inserting PINTs into unsuitable locations can have adverse effects on the listener. Werner et al. (2022) found that emulating a human-like PINTs pattern requires a deeper understanding of location optionality, duration variability, and the inclusion of breath noises. Werner et al. (2022) adds that in order to successfully model PINTs across language, the overly simple punctuation-dependent heuristics need to be expanded.

Speech synthesis systems have begun to reach human levels of naturalness when trained on read speech. The high level of naturalness has led to a variety of potential uses for PINTs in TTS. As the general segmental quality of TTS systems has improved, the focus has shifted towards suprasegmental elements, with the intent of creating natural and expressive sounding speech. For example, the inclusion of breath noises in speech synthesis may improve the naturalness and expressiveness desired in audiobooks, conversational assistants, and characters for movies and games. Robotics also has potential benefits from the implementation of PINTs in TTS, where the robot is able to employ appropriate PINTs while interacting with humans (Carlmeyer et al., 2018). However, PINTs still remain largely unexplored (Dall et al., 2016; Székely et al., 2019a) in the modeling and synthesis of spontaneous speech, with few attempts at modeling disfluent speech (Adell et al., 2007, 2008, 2010).

Another important aspect to consider is how the synthesis technique influences PINTs. For example, when comparing a parametric and unit-selection TTS systems, Aylett et al. (2020) found that personality is dependent on the type of TTS. The parametric voice was evaluated as less neurotic, and the unit-selection voice was evaluated as more open. Andersson et al. (2010) incorporated both FPs and lexical fillers into a unit-selection synthesis system and, via a perceptual study, found that their inclusion made the synthesis system more conversational without harming naturalness. Adell et al. (2007, 2012) found similar results that FP insertion did not reduce naturalness, in fact, listeners reported that the version without FPs was less natural. Using a deep neural network TTS system, Székely et al. (2017) were able generate FPs and modify the length of individual syllables for use in perceptual studies. They found that FPs and lengthenings could be used to modify the certainty or uncertainty of the generated speech. One of the difficulties in modeling PINTs for speech synthesis is the lack of consistent annotation methods. Difficulties arise due to annotator idiosyncrasies, different conversational tasks, and microphone conditions (Trouvain & Truong, 2012). Another important consideration during TTS training is the type of training data used. Gustafson et al. (2021) has shown it is beneficial to train a TTS system with different speaking styles (e.g., read and spontaneous), thereby the voice can generate a dynamic continuum of features present throughout the data.

The implementation of PINTs into TTS can further explore the relationship between listen-oriented benefits and PINTs. For example, FPs can improve TTS by reducing the cognitive load for the listener (Dall et al., 2016). It isn't always clear if phenomena that are beneficial to natural speech will be beneficial to synthetic speech as well. For example, Dall et al. (2014b) found that FPs helped reaction time in natural speech, but were detrimental to reaction time for synthetic speech. Betz et al. (2015) found that silences could be inserted into speech synthesis systems and still maintain acceptable quality. However, this finding was not replicated for FPs. Therefore, a primary goal was to evaluate the perceptual effects of PINTs in both natural and synthetic speech.

2.8 Recall and Education

In this dissertation, the terms 'recall' and 'recollection' are used interchangeably to refer to the ability to retrieve key information from memory. Research has shown that PINTs have listener-oriented benefits in both natural and synthetic speech. In natural speech, FPs have been found to improve the recall of story plot points (Fraundorf & Watson, 2011) and the following word (Corley et al., 2007), while silent pauses have been found to improve the recall of the following word (MacGregor et al., 2010). In synthetic speech, FPs can reduce the cognitive load for the listener (Dall et al., 2016), silences can improve digit recollection (Elmers et al., 2021a), and breath noises can aid in sentence recollection (Elmers et al., 2021b). These findings collectively illustrate that PINTs can improve the recall for small contexts, such as words or sentences in laboratory settings.

While the perceptual effects of PINTs have been researched in small contexts, larger contexts, such as education settings, have not been thoroughly evaluated. For example, Blau (1990) found that in a variety of proficiency levels, pauses were more beneficial for comprehension than using a "normal" speech rate or a mechanically slowed speech rate. Flowerdew & Tauroza (1995) found L2 subjects understood lecture material better when discourse markers were included, rather than excluded. When evaluating note-taking practices in lectures, Ewer (1974) found that the lecture must be as genuine as possible, which involves not imitating a 'reading' style. In other words, it is important not to omit elements of a 'lecture' style, such as PINTs. Similarly, Flowerdew & Miller (1997) espoused the need for providing students with genuine lecture material, otherwise they are not adequately prepared for authentic lecture settings. Mizuno (1990) found silences to be beneficial in a digit recall experiment conducted in a foreign language. Moniz et al. (2014) found that teachers use an intricate series of PINTs to allow for more time to edit their content during lectures, showing a different style of PINTs usage compared to dialogues.

The perceptual effects of PINTs can manifest differently for native speakers (NSs) and non-native speakers (NNSs). For example, Fayer & Krasinski (1987) found hesitations to be a major hurdle for NNSs to understand second language speech. Voss (1979) found similar effects with nearly 1/3 of all perceptual errors associated with hesitations, claiming that listeners sometimes mistake hesitations for lexical items or part of words. van Os et al. (2020) found for both NSs and NNSs that individuals who spoke quickly were rated as more fluent than individuals who spoke slowly. For NSs, answering a question too quickly or including long silences resulted in lower fluency ratings. However, for NNSs, only answers with long silences resulted in lower fluency ratings. van Os et al. (2020) shows that, as listeners, we perceive and evaluate NSs and NNSs differently, which is important to consider in educational settings. For example, silence duration has been shown to influence foreign accentedness in adult L2 speakers (Trofimovich & Baker, 2006; Kang, 2010). Specifically, Kang (2010) found that international teaching assistants who used shorter silences were evaluated as more native-like in terms of accentedness. Using FPs in Japanese to evaluate the prediction of difficult material, Watanabe et al. (2008) found that FPs did not influence the response time for low-proficiency NNSs, indicating that language proficiency is involved in the benefits of FPs. Rose (2017) found silences at clause boundaries were longer in L1 speech than L2 speech. Rose (2013) found for Japanese and English, that silence rate and duration were related to L1 performance, and that as their L2 skills improved their silence usage emulated their L1 silence patterns. In educational settings, NNSs are especially susceptible to difficulties when attempting to comprehend materials, such as university lectures. With many universities comprised of NNSs, it's important to understand when PINTs can help with recollection and when they disrupt the understanding key information.

The following chapters explore the influence of synthetic PINTs on recall, the detection and inclusion of PINTs in speech synthesis, and the implementation of synthesized PINTs in lecture-based scenarios.

2.9 PINTs Recall in Laboratory Setting

Chapter 3 investigates the effects of PINTs in speech synthesis on recall in small-context environments. This chapter consists of two experiments from Elmers et al. (2021a) and Elmers et al. (2021b).

In chapter 3.1 we evaluated the effects of an inserted silence on the recollection of numbers using synthesized speech. Participants heard segments that included a random 7-digit number. A silence of either 0 ms, 200 ms, or 500 ms was inserted prior to one of the digits. After listening to the 7-digit number, participants were tasked to fill in three missing consecutive digits. The results showed that participants' recall was improved for the number immediately after the silence. This effect was found only for the longer silence condition (500 ms) and not for the shorter silence condition (200 ms). Participants' response time was also higher when a silence was included in the stimuli.

In chapter 3.2, we conducted a partial replication of Whalen et al. (1995), which investigated whether inhalation breath noises in synthesized speech affected the recall of sentences. Whalen found that when inhalations were inserted at the beginning of a sentence that recall was improved over sentences that did not include an inhalation. Whalen used a formant synthesizer while we used a concatenative synthesizer. We investigated three conditions of breath noises: no inhalation (0 ms), short inhalations (300 ms), and long inhalation (600 ms). Our study was able to replicate the main findings of Whalen, that inhalations improved recall of the following sentence. However, the beneficial recall effect was only found for the long inhalation condition in our data. We also found that shorter sentences were recalled better than longer sentences.

Collectively, these experiments indicate that PINTs, namely silence and inhalation breath noises, can improve recollection for synthesized speech in single-sentence laboratory settings. These findings are similar to studies that evaluated the improvement of PINTs on recall in natural speech. These experiments also indicated that duration is an important aspect to the recall effect of PINTs, since only the long duration conditions improved recall. Both of these experiments used small-context stimuli and had limited PINTs control due to the systems used. Therefore, our next step was to develop a speech synthesis system that provided greater PINTs control and could be used to evaluate the effect of PINTs on recall in larger contexts.

2.10 PINTs Detection and Synthesis Generation

Chapter 4 evaluated different methods for detecting PINTs and the development of a custom speech synthesis system that produced PINTs with greater control. This chapter consists of two experiments from Elmers (2022) and Elmers et al. (2023).

Since PINTs are often omitted in corpora annotation, we investigated automatic approaches to annotating PINTs for large quantities of data. Chapter 4.1 compared the classification of PINTs using three different machine learning architectures: 1) a general neural network, 2) a convolutional neural network, and 3) a recurrent neural network. Mel-frequency cepstral coefficients were chosen as the input. An equal number of hyperparameters, numbers of layers, and neurons per layer was used to put a spotlight on the architectural differences of the models. Our initial hypothesis was that the recurrent neural network would outperform the other models since it is best able to handle temporal information, which is important considering that PINTs often co-occur. However, the three models performed similar to one another. All models successfully classified silences and breath noises, but were unable to successfully classify filler particles and tongue clicks, indicating that modeling PINTs simultaneously doesn't necessarily improve the accuracy for nearby PINTs. Overall, these results indicated that the annotation quantity and quality were better predictors of model accuracy than the model's architecture.

In chapter 4.2 we developed two text-to-speech (TTS) synthesis systems: 1) ControlledPINT, which incorporated labeled PINTs in the training data, and 2) AutoPINT, which did not include labeled PINTs in the training data. Both models were able to successfully generate a variety of PINTs but produced fewer PINTs and had a shorter total PINTs duration compared to natural speech. Moreover, the ControlledPINT version produced a greater number of PINTs and a longer total PINTs duration than the AutoPINT model. We conducted a perceptual experiment using the ControlledPINT model to evaluate the perception of certainty for the generated material. This experiment included four conditions: 1) a "fluent" condition that did not include PINTs, a long silence condition, a filler particle condition, and a combinatory condition that included a silence, filler particle "um", a tongue click, and an inhalation. The "fluent" condition without PINTs was rated as significantly more certain than the conditions that included PINTs, indicating that TTS certainty can be altered with the inclusion of PINTs material. The three PINTs conditions performed similarly, but the long silence condition was rated slightly higher than the filler particle condition, which was rated slightly higher than the combinatory condition.

Together, these two experiments indicated that it possible to detect and model PINTs for speech synthesis systems. The first experiment improved our understanding for automatically classifying multiple PINTs simultaneously, and improved the general understanding for classifying PINTs in the German language. Since the first experiment did not produce a model that could accurately detect all the PINTs material of interest, we developed an internal PINTs annotation system to train our TTS models. The second experiment provides a technological contribution with our models being, to the best of our knowledge, the first TTS systems to produce tongue clicks. Additionally, this study showcases that stimuli generated by TTS systems are a promising alternative to evaluate research questions in speech science. Since life-like communication is a continual goal of TTS systems, it is important to understand how PINTs and other spontaneous speech phenomena influence perception.

2.11 PINTs Recall in Lecture Setting

Chapter 5 synergizes the goals from the first two parts of this work, which involved investigating the recall effect of PINTs in synthetic speech and the creation of TTS systems that can generate PINTs. While chapter 3 focused on the recall effect in single-sentence laboratory settings, chapter 5 investigates larger contexts, specifically the influence of PINTs on the recall of synthetic lectures. This chapter incorporates three experiments. Chapter 5 is based on the publications Elmers & Trouvain (2022), Elmers (2023), and Elmers & Székely (2023).

In chapter 5.1, we developed a baseline PINTs usage by comparing real-world lectures from Yale University to the TOEFL iBT, a popular English-language proficiency exam used for university entrance. Annotations included 5 hours of material from 5 different Yale lectures (1 hour per speaker) and 15 minutes from the TOEFL lecture listening section. Since the Yale lectures were recorded during a three month semester, an additional analysis was made to compare intra-speaker PINTs variation at the beginning, middle, and end of the semester. Overall, PINTs filled 30% of the total lecture time for the Yale lectures, and 20% of the TOEFL lecture listening section. For the Yale data, speakers displayed unique PINTs usage with respect to count, duration, and frequency. Speakers were consistent with their PINTs usage, indicated by minimal intra-lecturer variation throughout the semester. Considering that PINTs made up 1/3 of real-world lecture time, it is important to understand how PINTs material influences the recall of key lecture information.

Most work evaluating the influence of PINTs on recall in both natural and synthetic speech has been conducted using single-sentence contexts. Chapter 5.2 established a baseline by evaluating the influence of PINTs on the recall of key information in lectures in natural speech. Participants included 45 native English listeners and 45 non-native L1 German listeners who heard English-language lecture segments, and answered content-based questions. Three conditions were evaluated: 1) an unmanipulated version, 2) a “silence” version, and 3) a “no PINTs” version where all PINTs material was removed. In the unmanipulated and “silence” conditions, half the critical information followed PINTs material. Overall, the recall of key information was reduced when the material was preceded by PINTs material. The listener’s first language did not have a significant effect on recall. However, native English listeners performed better during the “no PINTs” version, while the non-native L1 German listeners performed better during the original condition. This study was unable to find the recall benefit of PINTs in natural speech found in single-sentence laboratory experiments.

Chapter 5.3 replicated the experimental methodology used in chapter 5.2, but used synthesized speech rather than natural speech. The neural speech synthesizer from 4.2 that was trained on labeled PINTs annotations was used to generate experimental stimuli. This experiment also included three conditions: 1) a base version, 2) a “silence” version, where all non-silence PINTs were replaced with silence of the same duration, and 3) a “no PINTs” version where all PINTs material was removed. Half of participants were told they would hear computer-generated audio, while the other half were told that the audio was recorded with a poor-quality microphone. Participants also provided their subjective evaluations, such as interest level, via a questionnaire. Again, recall of key information was lower when the material was preceded by PINTs. Participants’ recall was not influenced by their first language or the condition they heard. Higher levels of interest resulted in a significant positive effect on the recall of key information.

This final section highlights multiple important points. First, PINTs are widely present both in real-world lectures and in the lecture listening section of the popular TOEFL iBT English-language proficiency test. Second, in both lecture recall experiments, the presence of PINTs lowered the recall of key information for both natural and synthetic speech. This finding contrasts with the results of single-sentence laboratory experiments, indicating that additional work is required to tease apart the influence of PINTs on recall. Finally, we were able to show that PINTs, in both natural and synthetic speech experiments, did not negatively influence non-native listeners more than native listeners.

Chapter 3

PINTs Recall in Laboratory Setting

3.1 Evaluating the Effect of Pauses on Number Recollection in Synthesized Speech

3.1.1 Abstract

This study investigates the effects of an inserted pause on digit recollection for synthesized speech. Participants took part in a perception experiment which involved listening to a 7-digit random number that was rendered by a speech synthesis system. Some of the stimuli had pauses (200 ms or 500 ms in duration) inserted before one of the digits, while others did not include a pause. Immediately following each stimulus the participants were asked to provide a missing sequence of three adjacent digits. Results indicate that recall accuracy is improved immediately following a pause. Additionally, we found a significant effect for a pause duration of 500 ms but not for a pause duration of 200 ms. When investigating response time, we found that participants' response time increased when a pause was present. Overall, the results show that pauses have a role to play in synthesized speech. This research can be regarded in the context of investigating pauses and pause-internal particles (e.g. breath noises) in synthesized speech and the effects they have for human listeners.

3.1.2 Introduction

Speech synthesis systems have become ubiquitous in the banking and telephone industries. This progression has created situations where the average person is required to interact with synthesized recordings, often of strings of numbers (e.g. credit card and bank account). These exchanges are regularly complicated by a necessity for high accuracy and show no redundancy, in contrast to most other types of linguistic information. There is evidence that telephone numbers are grouped prosodically, which helps to recall those numbers (Baumann & Trouvain, 2001). This prosodic grouping is usually realized by rhythmic features such as alternations of accented and unaccented digits within a minor prosodic phrase. The boundaries of these minor prosodic phrases are sometimes marked by a short pause. Therefore, in this work, a perceptual experiment was conducted to investigate the effects of the presence of a pause on recollection accuracy for synthesized digits. More specifically, this experiment endeavoured to investigate the consequences of pauses on short-term digit recollection.

We previously conducted a pilot study with a similar focus using MaryTTS (Schröder & Trouvain, 2003). The pilot study focused on a single, short pause duration of approximately 200 ms and compared it against a no-pause condition. The results from the pilot study indicated that pauses in synthesized speech could improve digit recollection. For the current study, we aimed to elaborate on our previous findings and document the improvement of digit recollection with pause insertion for synthesized speech.

When researching TTS systems for the present study, multiple systems were considered, including MaryTTS (Schröder & Trouvain, 2003), Festival (Taylor et al., 1998), and Amazon Polly (2016). Interestingly, none of these systems created pauses automatically when generating synthesized digit sequences, and they all required some form of text markup. Both MaryTTS and Festival occasionally experienced problems where part of the digit audio was truncated. To avoid any fractured audio we opted to use Polly to synthesize the audio clips. While all three TTS systems use voices created by concatenative synthesis, Polly was found to be superior in audio quality when compared to MaryTTS and Festival. This punctuated our decision to move forward using Polly and facilitated our intention to keep any audio quality discomfort as low as possible when listening to the audio clips. This in turn, allowed the participants to focus on the prompted digits rather than audio irregularities.

Another change that we made between the pilot and the present experiment was the addition of a second pause duration. In the pilot study, we observed a tendency towards the pause condition in the recollection accuracy between no pause and the 200 ms pause insertion. In this study we have added an additional pause duration of 500 ms. Our decisions regarding pause duration were based on a large multilingual study of silent pause durations (Campione & Véronis, 2002). By adding the longer pause to the experiment we hoped to see further exaggerations of the results indicated in the pilot study.

3.1.3 Method

Material

Participants listened to audio clips of synthesized speech that contained a randomized 7-digit number (e.g. 3852791). A 7-digit number was selected based upon Miller's Law

(Miller, 1956), which states that the average person's short-term memory capacity is 7 +/- 2. Participants were asked to type a 3-digit sequence. There were 5 potential sequences:

1. {1 2 3} 4 5 6 7
2. 1 {2 3 4} 5 6 7
3. 1 2 {3 4 5} 6 7
4. 1 2 3 {4 5 6} 7
5. 1 2 3 4 {5 6 7}

In both pause conditions (200 and 500 ms), a pause was inserted prior to one of the digits participants were asked to type, i.e., the digits within the curly brackets. A 3-digit graphical sequence was chosen to mask the critical digit, viz. the digit following the pause. The experiment included three pause durations: 0 ms, 200 ms, 500 ms. The durations of 200 and 500 ms were chosen to represent a short and a normal pause, respectively. The first and last digits were included as a baseline to confirm primacy and recency effects (McLeod, 2008).

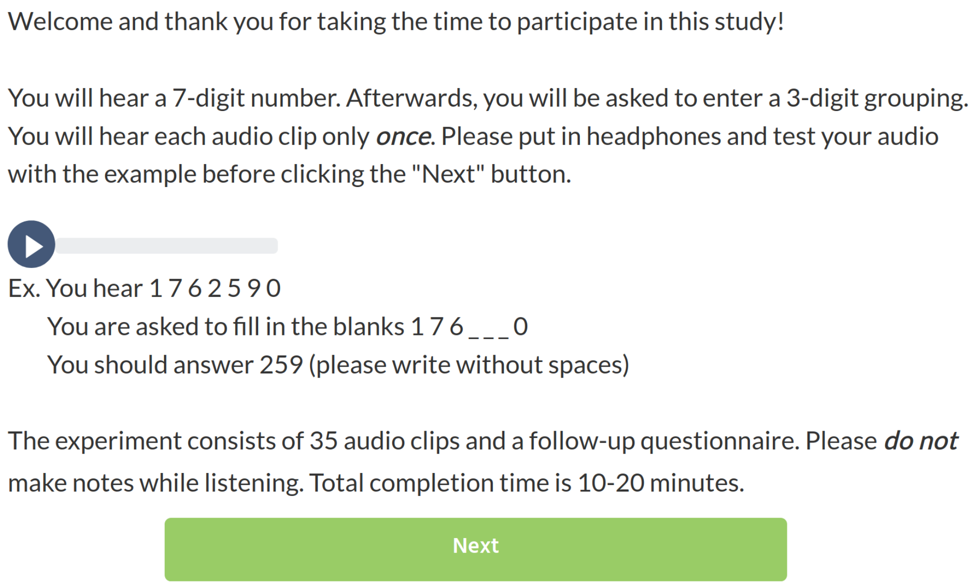
The stimuli were generated using Amazon Polly's TTS service with Joanna's voice, the standard TTS voice generated using concatenative synthesis. The pauses were inserted by using an instruction in the Speech Synthesis Markup Language (SSML) (Baggia et al., 2010) indicating the pause duration in milliseconds.

Experiment

The material was uploaded to the online experiment platform Labvanced[1] (Finger et al., 2017). When beginning the experiment, participants were asked to use headphones and test their audio on the instruction screen. Then participants were instructed that they would hear a sound clip with a 7-digit number and that each clip would be played only once. After listening to the prompted clip, a box appeared on the screen and participants were asked to type a 3-digit sequence from the 7-digit number. Figure 3.1 shows the instructions participants received before beginning the experiment. And Figure 3.2 shows the screen that appeared after listening to the audio clip, where participants entered the 3-digit sequence.

The experiment consisted of 35 audio clips, which included two trial runs that were not counted in the results, and a follow-up questionnaire. The experiment was designed so that each participant would experience every condition, including all pause locations, sequences, and durations. Total completion time was 10–20 minutes for each subject.

[1] Accessed via https://www.labvanced.com/ on Dec. 01, 2021.



Welcome and thank you for taking the time to participate in this study!

You will hear a 7-digit number. Afterwards, you will be asked to enter a 3-digit grouping. You will hear each audio clip only *once*. Please put in headphones and test your audio with the example before clicking the "Next" button.

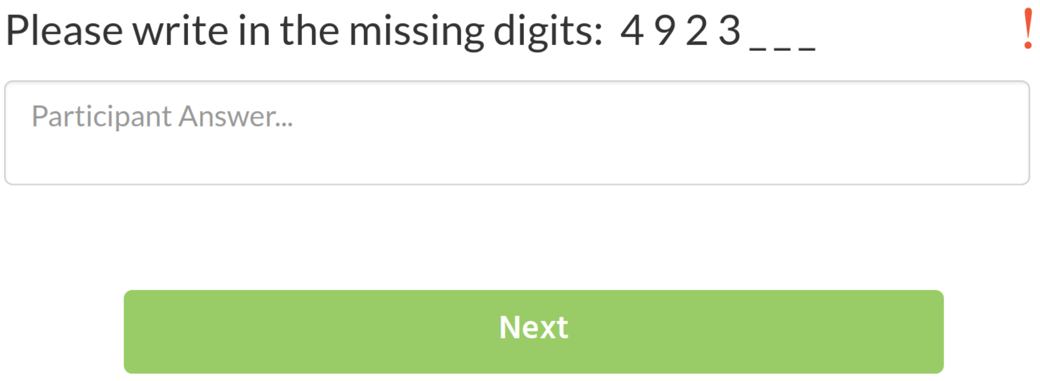
Ex. You hear 1 7 6 2 5 9 0

You are asked to fill in the blanks 1 7 6 _ _ _ 0

You should answer 259 (please write without spaces)

The experiment consists of 35 audio clips and a follow-up questionnaire. Please ***do not*** make notes while listening. Total completion time is 10-20 minutes.

Figure 3.1: The instructions participants received before beginning the experiment.



Please write in the missing digits: 4 9 2 3 _ _ _

Participant Answer...

Figure 3.2: The screen participants saw after listening to the stimulus. Here they would enter a 3-digit sequence.

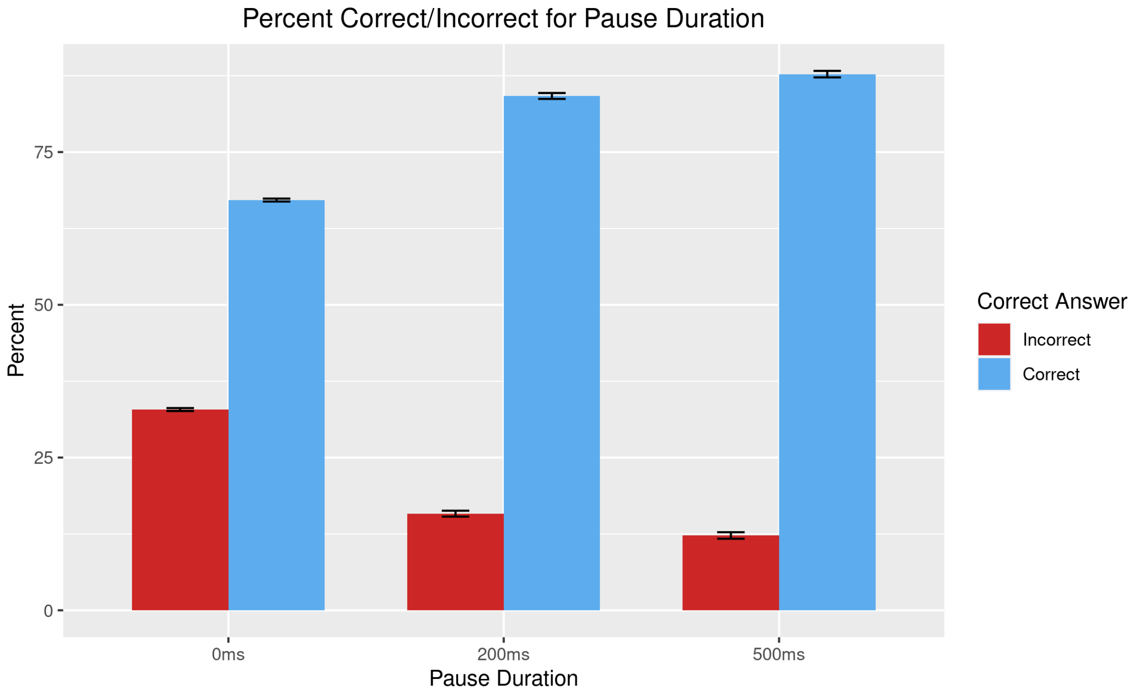


Figure 3.3: Accuracy of the critical digit following the pause, or the sequence-central number in the no-pause condition, arranged by pause duration.

Participants

Participants were recruited from an online service using Prolific[2] (2014) and were offered payment for their time. There were a total of 15 subjects (10 F and 5 M, age range 25–60, mean age 36.2 years). All participants, except one, self-reported no form of hearing impairment. The participant who self-reported hearing impairment was excluded from the analysis. In order to determine familiarity with synthesized speech, subjects were asked, 'how often do you listen to text-to-speech audio?' Possible responses included were, “never”, “monthly”, “weekly”, and “daily”. Of the 15 participants, 8 (53%) indicated that they never listen to TTS audio, 4 (27%) indicated monthly, 1 (7%) indicated weekly, and 2 (13%) indicated daily usage.

3.1.4 Results

The presence of a pause resulted in a higher recollection accuracy for the following digit than when the pause was absent. Similarly, Figure 3.3 shows that both the short (200 ms) and the normal (500 ms) pause durations caused a higher accuracy than the condition with no pause.

[2] Accessed via https://www.prolific.co/ on Dec. 01, 2021.

Table 3.1: Model 1: GLMM Results Accuracy~Pause Occurrence + (1 | Subject) + (1 | Item).

	estimate	std. Error	z	p
$(Intercept)$	0.8947	0.6915	1.294	0.1957
$PauseOccurs$	1.6214	0.7475	2.169	< 0.05

Accuracy Modeling

Multiple statistical models were analyzed for accuracy of the critical digit, i.e. the digit following the pause, and for response time (RT). The response variables were analyzed by using generalized linear mixed-effects models (GLMMs) from the lme4 package (Bates et al., 2015) in R (R Core Team, 2021).

Model decisions were made bottom-up, beginning with only random intercepts for *subject* and *item*, and progressively adding fixed effects. Random slopes for the fixed effects were added for *subject* assuming no issues from over-fitting or non-convergence. Models were compared via the Akaike information criterion (AIC) (Akaike, 1973) to determine unexplained variance. If the AIC decreased by at least two points, then a factor was kept in the model.

The GLMMs were analyzed with *accuracy* of the critical digit (binary categorical variable, 0 for incorrect and 1 for correct) as the response variable. Models with the following predictor variables were evaluated: *pause occurrence* (binary categorical variable: 0 for absent, 1 for present), *pause duration* (factor with three levels: 0 ms, 200 ms, and 500 ms), *sequencing* (factor with 5 levels), and *digit position* (factor with 6 levels). For digit position, the first digit was not taken into account as it was never the critical digit. Due to collinearity effects, pause occurrence and pause duration were modeled separately.

Model 1 (Table 3.1), the model with the lowest AIC, included pause occurrence as the only fixed effect. Subject and item were included as random intercepts. The GLMM used a binomial family and logit link. This model shows that the presence of a pause is statistically significant and increases recollection accuracy (estimate (log-odds) = 1.6214, SE = 0.7475, z = 2.169, $p < 0.05$).

Model 2 (Table 3.2), the model with the lowest AIC, included pause duration as the only fixed effect. Subject and item were included as random intercepts. The GLMM used a binomial family and logit link. This model indicates that the pause duration of 500 ms was statistically significant (estimate (log-odds) = 1.9911, SE = 0.8309, z = 2.396, $p < 0.05$) and is beneficial for recollection accuracy. However, the 200 ms pause duration was not statistically significant. Models were also analyzed for accuracy predicted by RT, yet none of the models achieved a lower AIC than the model with only random effects.

Table 3.2: Model 2: GLMM Results Accuracy~Pause Duration + (1 | Subject) + (1 | Item).

	estimate	std. Error	z	p
$(Intercept)$	0.8940	0.6871	1.301	0.1932
$PauseDur200ms$	1.3019	0.7918	1.644	0.1001
$PauseDur500ms$	1.9911	0.8309	2.396	< 0.05

Table 3.3: Model 3: GLMM Results RT~Pause Occurrence + (1 | Subject) + (1 | Item).

	estimate	std. Error	z	p
$(Intercept)$	4930.62	26.48	186.19	< 0.001
$PauseOccurs$	363.40	28.09	12.94	< 0.001

Response Time Modeling

The subject's response time (RT) was also recorded. Participants were only able to hear the clip once, and the RT timer started as soon as the audio clip ended. Upon submitting their answers the RT timer finished. The participants' RT had a highly positive skew, therefore, values that exceeded 3 standard deviations above the mean were excluded. Even with these values removed RT still skewed positive but was far less extreme. Even so, a gamma distribution was chosen. Additionally, while investigating RT, only correct answers were included in the models.

For Model 3 (Table 3.3), the model with the lowest AIC, included only pause occurrence as a fixed effect. Subject and item were included as random intercepts. A GLMM, with a gamma family and identity link, was chosen over a LMEM with a log-transformation of RT. This decision was made to prevent issues that can occur from seeking normality of a log-transformed RT (Lo & Andrews, 2015).

Table 3.3 shows that pause occurrence is significant for RT (estimate = 363.40, SE = 28.09, z = 12.94, $p < .001$). Interestingly, the effect is an increase in RT. The coefficient value of 363.40 is similar to the average duration between the two pause durations, 200 and 500 ms. This duration might be representative of an abstract pause involved when the participants mentally recall the synthesized digits, before typing their answer. Models were also analyzed for RT predicted by pause duration, yet none of the models achieved a lower AIC than the model with only random effects.

3.1.5 Discussion and Summary

In this study, participants were tasked with listening to a 7-digit clip of synthesized speech to determine if a pause affected their recollection accuracy for the following digit. This study aimed at improving an effect found in our pilot study, specifically that a pause in synthesized speech aided in digit recollection. This study made improvements over the pilot by including: a higher quality concatenative TTS system, an additional 500 ms pause duration, and investigating RT. Using GLMM models, we have shown, generally, that the presence of a pause indeed affects recollection accuracy. Moreover, we also found that the 500 ms pause duration improved digit recollection. However, we were unable to confirm the results from our pilot study that a 200 ms pause duration improved digit recollection. These results emphasize the importance of further research on pauses in synthesized speech.

An important aspect of synthesized digit sequences is the prosodic structure, specifically how the number sequences are grouped and the number of groups. All stimuli in this study contained two prosodic groups. The first included all digits up to the pause, while the second consisted of all the digits following the pause. It is important to investigate these prosodic structures with more attention. Additionally, in the future we could include basic grouping strategies (e.g. 3-2-2) for 7-digit numbers (Baumann & Trouvain, 2001) to evaluate different prosodic groups and their influence on digit recollection accuracy.

In the current study we have shown that the presence of a pause also influences response time, with RT increasing when a pause is present. Results indicate that the participant does not differentiate between pause durations while recalling the digits. The RT model showed that participants might be retaining some abstract pause duration in their mind during recollection. RT was measured after the sound clip finished and without a delay. Future research should evaluate whether the duration between when the clip finishes, and when the participant is able to respond, affects their accuracy. A promising next step in this research would be to investigate pause-internal particles, such as breath noises, for their effects on synthesized speech digit recollection.

3.2 Take a Breath: Respiratory Sounds Improve Recollection in Synthetic Speech

3.2.1 Abstract

This study revisits Whalen et al. (1995) by evaluating English speaking participants in a perception experiment to determine if their recollection is affected by including breath noises in sentences generated by a speech synthesis system. Whalen found an improvement in recollection for sentences that were preceded by a breath noise compared to sentences without one. While Whalen and colleagues used formant synthesis to render the English sentences, we use a modern concatenative synthesis system. The present study uses inhalations of three different lengths: 0 ms (no breath noise), 300 ms (short breath noise), and 600 ms (long breath noise). Our results are consistent with Whalen and colleagues for the 600 ms condition, but not for the 300 ms condition, indicating that not all inhalations improved recollection. The present study also found a significant effect for sentence length, illustrating that shorter sentences have higher accuracy for recollection than longer sentences. Overall, the present study indicates that respiratory sounds are important to the recollection of synthesized speech and that researchers should focus on longer and more complex types of speech, such as paragraphs or dialogues, for future studies.

3.2.2 Introduction

In the present study, we examined pause particles in synthesized speech. Previous work by Whalen et al. (1995) (henceforth Whalen) found that English speaking participants' recollection, sometimes referred to as recall (as in Whalen), was better for sentences preceded by a breath noise than those not preceded by a breath noise. Whalen's study was conducted using a formant synthesizer, KLATTALK (Klatt, 1982). In contrast, Trouvain & Möbius (2013) used concatenative synthesis to evaluate the perception of telephone numbers preceded by an inhalation. They found that the majority of subjects did not have a preference. The results from Whalen et al. (1995) and Trouvain & Möbius (2013) offer conflicting interpretations of the effect of breath noises in synthesized speech, which called for further investigation. Elmers et al. (2021a) found that the insertion of a silent pause in a 7-digit sequence improved the recollection of the following digit. An appropriate next step was to evaluate breath noises and revisit Whalen's study.

The primary objective for this experiment is to clarify the conflicting interpretations between Whalen et al. (1995) and Trouvain & Möbius (2013). Therefore, we endeavoured to examine if breath noises aid in recollection. In an effort to investigate this question, the present study closely mirrors the Whalen study, with some updates concerning technology. Specifically, we used Amazon Polly (2016) to generate our stimuli and a web-based platform to conduct the experiment. By including these modifications, and other nuanced updates, we intend to contribute research to pauses and pause particles in synthesized speech.

3.2.3 Method

Comparison of the Present Study and Whalen

The present study is a partial replication of Whalen, combining ideas from their experiments 1, 3 and 4. In each experiment participants listened to synthesized audio and, afterwards, wrote down what they heard. Experiment 1 focused on the effect between a breath noise and a no breath noise condition, with each condition separated into a single block. For example, the participants would hear a block of 20 sentences each preceded by a breath noise, followed by a second block of 20 sentences not preceded by a breath noise. The opposite ordering of blocks was also included. They found a significant effect for breath noises on the improvement of recollection. Moreover, the no breath noise condition did not have a significant effect on recollection improvement. Lastly, they found an improvement due to practice.

Experiment 3 and 4 maintained the breath noise/no breath noise conditions from experiment 1, but with more specificity. In experiment 3, rather than using the same block system from experiment 1, the breath noises were inserted randomly before sentences. Once again practice was found to be significant, but breath noises were not significant. While experiment 3 focused on random distribution of the breath noises, experiment 4 focused on appropriateness. In their earlier experiments they had maintained the appropriateness of the breath noise. In other words, short sentences were only preceded by the short (mean duration ∼600 ms) breath noise and long sentences were only preceded by the long (mean duration ∼740 ms) breath noise. In experiment 4, they tested appropriateness in a way that both short and long breath noises appeared before both short and long sentences. They found appropriateness was not significant but they indicated this may be due to the small range of sentence lengths.

The present experiment synergizes many of the aforementioned ideas from Whalen. We incorporated the breath noise vs no breath noise conditions from experiment 1. We assigned breath noises randomly before sentences, rather than in blocks (like experiment 3). Lastly, we evaluated appropriateness by inserting short breath noises before long sentences, and vice versa (like experiment 4). This experiment examines the following durational conditions: a 0 ms no breath noise (henceforth NO-brn), a 300 ms breath noise (henceforth SHORT-brn), and a 600 ms breath noise (henceforth LONG-brn). Table 3.4 contains a comparison between the present study and Whalen for breath noise durations and sentence lengths. Participants in both experiments heard synthesized audio and recollected what they heard. However, in the present experiment participants typed their responses after each stimulus rather than writing

Table 3.4: Mean (SD) for breath and sentence lengths reported here compared to Whalen (SD was not reported in Whalen).

	present study	**Whalen**
Short breath duration (in ms)	300	597
Long breath duration (in ms)	600	738
Short sentence length (in words)	8.5 (2.0)	8.1
Long sentence length (in words)	16.2 (3.6)	15.2

them by hand. The experimental design in Whalen is easily converted into a web-based study like we did here.

Creating the Stimuli

For this experiment we used Amazon Polly, which Amazon describes as a "Text-to-Speech service that uses advanced deep learning technologies to synthesize speech that sounds like a human voice" (Amazon Web Services, 2016). The documentation for Polly does not provide further information beyond "advanced deep learning technologies" to clarify how the breath noises were created or the amount of breath noise variation. Polly's breath feature announcement claims that Polly can parrot the sounds of both inhalation and exhalation for normal speech. However, in our time working with Polly only inhalations could be identified. Additionally, the breath tags required for synthesizing respiratory sounds are currently only available for the standard voices, which use concatenative synthesis, not for the neural voices.

Polly includes an automated mode which allows the user to indicate (using preset values) the volume, frequency and duration for the synthesized breaths. The current experiment uses the manual mode to specify exact locations, and to customize the duration and volume. The breath noises (for both automated and manual mode) must be indicated using text mark-up, specifically Speech Synthesis Markup Language (SSML) (Baggia et al., 2010).

Whalen's stimuli were created with KLATTALK (Klatt, 1982), a formant synthesizer. Their breath noises were made from recordings of a person with a similar vocal tract to the voice model of their synthesizer. They recorded a total of six breath noises (3 short and 3 long) to add variety and factor out any oddities. Additionally, they indicated that their sentences were *not* completely comprehensible, but every sentence was answered correctly by at least one participant.

With the goal of creating more natural and expressive speech, we used Polly to generate inhalation sounds. The three conditions for this experiment were: 1) NO-brn (i.e. 0 ms), SHORT-brn (mean duration: 300 ms), and LONG-brn (mean duration: 600 ms). Our justification for the short and long inhalation durations are from a study on phrase-initial inhalation noises (Werner et al., 2021), which differ from phrase-

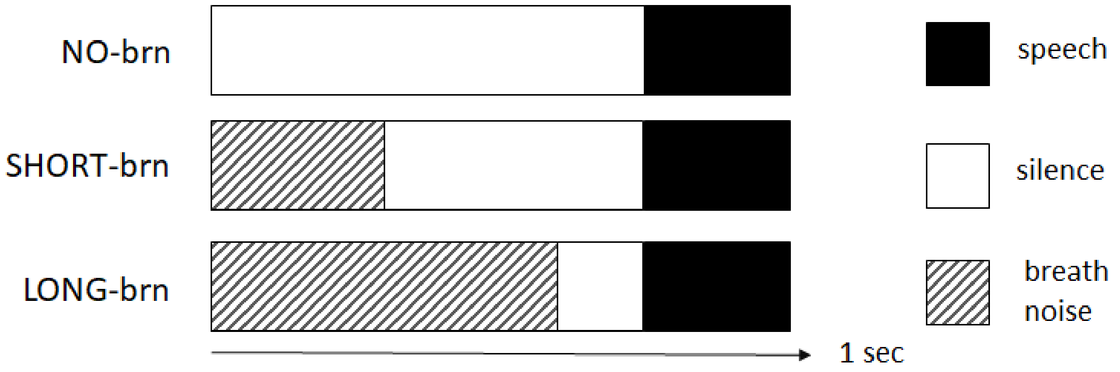


Figure 3.4: Schematic for the first second of the stimuli in the three conditions.

internal inhalations.

We chose to use Polly's "default" breath noise since it is ∼300 ms in duration. We also chose Polly's "default" intensity value since it is ∼40 dB, which is consistent with what (Werner et al., 2021) found as a median intensity for phrase initial inhalations. Werner et al. (2021) also found a median value of ∼140 ms for the right-edge pause between the phrase-initial inhalation and the onset of speech. Polly naturally inserts a ∼50 ms right-edge pause between the inhalation and the speech, so we increased this to a total of 150 ms by adding an additional 100 ms of silence (Fig. 3.4).

The present study uses a total of 28 different sentences (24 experimental, 3 practice, and 1 for instruction). For the 24 experimental stimuli, 12 were short sentences (mean number of words = 8.5, SD = 2.0, range: 5–12) and 12 were long (mean number of words = 16.2, SD = 3.6, range: 13–26). Some sentences included simple numbers, but none included complex or alphanumeric expressions. The sentences were created with Polly using the aforementioned methodology and consisted of situations that are typically discussed with conversational assistants such as weather, schedule information, restaurant bookings, etc.

Three versions of each sentence were created using our three conditions (NO-brn, SHORT-brn, and LONG-brn), resulting in a total of 72 tokens. These 72 tokens were evenly divided into three lists, designed in such a way that each sentence appeared only once per list. Additionally, the lists were balanced to achieve an equal number of each breath noise condition. The tokens in each list were randomized, so that different lists had a different ordering of sentences. However, participants who saw the same list encountered the sentences in the same order.

Participants

We created our web-based experiment using Labvanced[3] (Finger et al., 2017) to present the audio stimuli to the participants and collect their typed answers, question-

[3] Accessed via https://www.labvanced.com/ on Mar. 03, 2021.

naire information, and response time (RT). Participants were recruited with Prolific[4] (2014) and consisted of 63 monolingual English participants (mean age 36.92 years; age range 18–70 years; 29 females, 33 males, 1 non-binary; 59 British accented, 2 American accented, 2 Australian accented) who were paid for their participation. One participant indicated hearing impairment and was excluded from the results. For the experiment, subjects were instructed to type what they heard exactly as they heard it. Subjects were presented with one of three lists. Each list consisted of the same 24 sentences. However, they varied in which breath noise condition (NO-brn, SHORT-brn, LONG-brn) preceded the sentence, and in the overall ordering of the stimuli. Participants listened to one audio clip during the instruction screen which was followed by three practice sentences (not included in the results). The practice sentences included examples that were preceded by a breath noise and some not preceded by a breath noise. After completing the listening portion they filled out a questionnaire.

Scoring and Data Processing

After collecting the participants' results, we standardized the data by tokenizing, removing punctuation and extra white space, converting words to lowercase, and correcting some spelling errors. For example, if a participant typed "appoxiately", we corrected it to "approximately", and counted it as correct during the scoring. However, homophones or words that did not preserve the intended meaning of the sentence were not corrected. For example, if a participant wrote "weight" instead of "wait" in the context of waiting for a table at a restaurant then their word was not corrected, and consequently, not scored positively.

After standardizing the data, participants were scored based on how many of the correct words they had included in their response. They were awarded 1 point for each correct word. In the present study we focused on whether the correct word was included, not on the order. Whalen's scoring method provided one point for a correct word in the correct location. A mostly correct word was worth 0.5. A correct word in the incorrect location provided 0.5. Whalen scored homophones as correct and did not encounter semantically related words. In the present study, the scoring system was simplified so that participants were awarded 1 point if the word in their submission was found in the canonical version (i.e., the correct version). The present study and Whalen, counted function and content words equally when scoring, since the TTS systems used in the two studies did not reduce function words as in human connected speech. Scores were normalized by dividing the participant's score by the length (i.e., number of words) of the canonical version of the sentence. Normalized scores ranged from 0 to 1. The differences in scoring methods might affect differences between the two studies. However, within the study, since all stimuli were scored using the same method, there is a level of consistency when comparing the scores.

[4] Accessed via https://www.prolific.co/ on Mar. 03, 2021.

Table 3.5: Scores normalized by number of words for different conditions.

condition	mean	sd
All Conditions	0.909	0.154
NO-brn	0.902	0.164
SHORT-brn	0.902	0.160
LONG-brn	0.923	0.136
Length Short	0.959	0.110
Length Long	0.860	0.175

3.2.4 Results

The mean and standard deviation for the different breath noise conditions can be seen in Table 3.5. When looking at the mean scores for all conditions, it is clear that participants are already scoring near the normalized score ceiling, which can also be seen in Fig. 3.5. When looking at the individual breath noise conditions, we find higher scores for the LONG-brn condition compared to the NO-brn and SHORT-brn conditions. As for length, we also find a score difference between short and long sentences.

Statistical models were analyzed with linear mixed-effects models (LMEM) from the lme4 (Bates et al., 2015) package (Version 1.1.25) and the lmerTest (Kuznetsova et al., 2017) package (Version 3.1.3) in R (R Core Team, 2021) (Version 3.6.3). Models were made using backwards selection, i.e., starting with the maximal model for fixed and random effects and gradually reducing (starting with random slopes) in the case of over-fitting or non-convergence. Models were compared with the Akaike information criterion (AIC) (Akaike, 1973), which calculates unexplained variance, and the model with the lowest AIC was considered as the model with the best fit.

The final model was: $lmer(NormalizedScore \sim BreathNoise + Length + (1 \mid Subject) + (1 \mid Sentence), REML = FALSE)$. This model includes breath noise duration and sentence length as fixed effects (without an interaction term). As random effects, intercepts were included for both the subject and the individual sentences. Visual inspection for the residual plot revealed deviations from homoscedasticity and a violation of normality (partly caused by the ceiling effects). However, Schielzeth et al. (2020) has shown that linear mixed-effects models are robust to these types of violations. Our analysis revealed a main effect for the LONG-brn condition ($Estimate$ = 0.02077, SE = 0.00722, t = 2.877, p < 0.01) and the short sentence length ($Estimate$ = 0.09894, SE = 0.02986, t = 3.314, p < 0.01). These main effects indicate an increase in recollection of the sentence. We found that shorter sentences are recalled better than longer sentences, and that sentences immediately preceded by a LONG-brn are recalled better than sentences preceded by the NO-brn or the SHORT-brn.

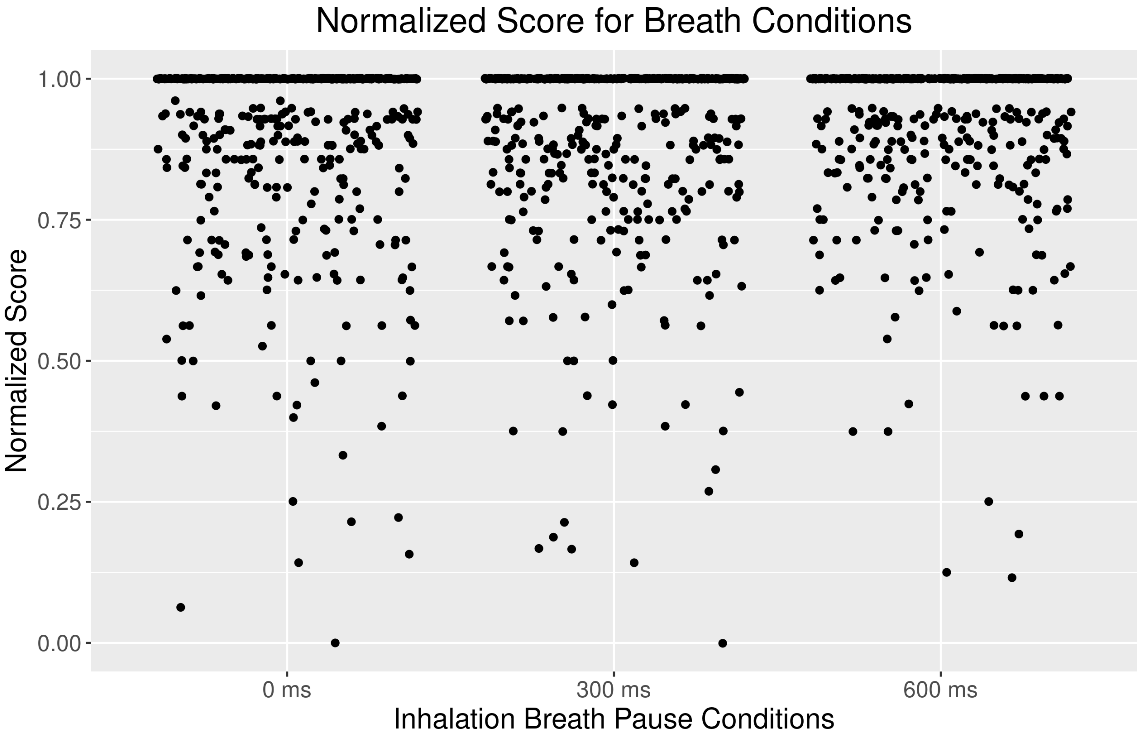


Figure 3.5: Scatterplot for score normalized by number of words for each of the breath noise conditions.

3.2.5 Discussion

The present study replicated one of the major findings from Whalen, namely that the LONG-brn condition improves recollection. With these results in mind, future research can investigate the following: duration, learning effects, sentence length, and measuring recollection.

Duration

When designing the experiment, the first author found the SHORT-brn to be most natural, while the LONG-brn appeared abnormally long. However, the SHORT-brn condition was not significant while the LONG-brn condition was significant. The short and long breath noises used by Whalen were longer than the versions used in the present study, and found to improve recollection. Importantly, the present study's LONG-brn was approximately the same duration as Whalen's short condition. This finding may indicate that exaggerated breath noises, and possibly other pause particles, are more suitable for synthesized speech with respect to recollection.

There are many hypotheses that could explain the improvement in recollection caused by various particles in speech (Fraundorf & Watson, 2011), including breath noises. While we describe these options, we do not position one as the primary rationale for recollection improvement. Three possibles hypotheses are: 1) processing-time hypothesis, i.e., the breath noises are providing more time for the listener to process what they hear, 2) attention orienting hypothesis, i.e., the breath noises are drawing the listener's focus, and 3) predictive processing hypothesis, i.e., participants use the breath noises to predict upcoming speech content. Future work should further investigate the specific mechanisms for improving recollection in synthesized speech.

Learning Effects

Whalen found that participants performed better during the second half of the stimuli than during the first half (i.e. learning effect). The present study did not find any kind of learning effect, possibly due to improvements in audio quality for modern TTS systems. Another possibility is that listeners have become more acclimated to hearing synthesized audio. In a follow-up questionnaire, participants were asked how often they listen to computer-generated audio, such as conversational assistants or in-car navigation. Only 11 of the 63 participants reported never listening to computer-generated speech; however, this number might be inaccurate if participants misunderstood potential situations in which they hear computer-generated audio, such as robocalls or online videos.

Sentence Length

Whalen measured sentence length in number of words. Consequently, the present study also measured length via number of words, in order to maintain parity with Whalen. Ideally, length would be evaluated using a more stable metric such as a speech timing unit, e.g., number of syllables. This would alleviate the problem that arises when two sentences share the same number of words but vary greatly in their number of syllables.

The present study found high recollection scores for short and long sentences. Therefore, future work should include longer material lengths, such as paragraphs or fragments of dialogue. In the present study, short sentences (mean length = 8.5 words) had a mean accuracy of 0.959, whereas the long sentences (mean length = 16.2 words) had a mean accuracy of 0.860. The high quality of the synthesizer allows participants to not only understand the material, but repeat it verbatim, with near perfect accuracy. While we see an accuracy drop in the longer sentences, future experiments should investigate both longer and more complex sentences and discourses. In fact, Braunschweiler & Chen (2013) concluded that paragraphs and longer sentences are important and might improve naturalness for the listener by reducing the monotony and improving the prosody of speech synthesis. Interesting examples would be paragraphs of material, such as audiobooks, or dialogic conversation between humans and conversational agents. Finally, it would be interesting to look into semantically unpredictable sentences to see if these results for recollection hold.

Measuring Recollection

Both the present study and Whalen tested the participants' ability to recollect the exact message they had heard. While typing or writing their answer, participants are required to focus on spelling, potentially reducing the amount of effort they can give to the general content. It is important to think about what metrics and constructs are used to measure participant recollection, since there are many different ways to measure understanding and memorization. One possible alternative could have participants listen to an audio clip and record a summary in their own words, similar to Fraundorf & Watson (2011), so that a participant's score would be dependent on overall comprehension rather than a word-for-word memorization. Another alternative could provide participants with multiple-choice questions. Future work should focus on a particular format to evaluate specific details with more nuance.

3.2.6 Conclusion

The present study investigated the effect of an inserted breath noise on recollection of synthesized speech, similar to Whalen et al. (1995). Our results are comparable to the results found by Whalen and colleagues. Three breath noise conditions were evaluated, a NO-brn (i.e. 0 ms) condition, a SHORT-brn (mean duration: 300 ms) condition, and a LONG-brn (mean duration: 600 ms) condition. Participants displayed a high level of recollection overall, even in the NO-brn condition. The LONG-brn improved recollection, whereas the SHORT-brn did not. We also found a significant effect for sentence length, which indicates that recollection is better in shorter sentences.

This experiment evaluated breath noises in single sentence contexts, avoiding connected speech due to difficulties in determining whether the breath noise influences the planning of the upcoming sentence or is a consequence of the preceding speech. Therefore, we chose to investigate breath noises in a smaller, more manageable context before looking towards longer and more complex forms of discourse in the future. Beyond investigating recollection abilities as a function of breath noises, future work should also view this phenomenon from the perspective of naturalness, which is important for maintaining expressiveness without sacrificing the pleasantness of synthetic speech.

Chapter 4

PINTs Detection and Synthesis Generation

The experiments in this chapter examined the following: 1) the automatic detection of PINTs, 2) the training of neural synthesis systems with pause material, and 3) the rendering of synthetic material that incorporated PINTs. The automatic detection and classification of PINTs is an important step for training TTS systems. Experiment 1 (chapter 4.1) investigated a variety of machine learning methods, with the goal of modeling multiple PINTs simultaneously. Experiment 2 (chapter 4.2) explored the training of two neural synthesis systems, one that generated PINTs via the insertion of designated labels, and a second system that rendered PINTs probabilistically. These experiments used speech signal processing and machine learning methods to model PINTs in both natural and synthetic speech. Overall, this chapter ties together the detection and rendering of PINTs material with speech technology.

4.1 Comparing Detection Methods for Pause-Internal Particles

4.1.1 Abstract

This study investigates different machine learning architectures for classifying pause-internal phonetic particles (PINTs), such as filler particles (FPs), breath noises complementary to silences, and tongue clicks. Many of these PINTs co-occur, and by modeling them simultaneously, the aim is to improve the classification accuracy for the surrounding PINTs as well. An annotated subset from a German spontaneous speech corpus was used for modeling. Mel-frequency cepstral coefficients were used as inputs to model PINTs with three kinds of neural networks: a general neural network, a convolutional neural network, and a recurrent neural network. The models used

the same hyperparameters, number of layers, and number of neurons for those layers, so that the focus was put onto the model architecture. The recurrent neural network was expected to perform the best since it is able to capture temporal information; however, all models performed similarly. The models performed best at classifying silent segments, followed by inhalations and exhalations. However, all models failed to accurately classify FPs and clicks, indicating that modeling PINTs simultaneously doesn't always improve accuracy for surrounding PINTs. These findings suggest that accurate classification is more dependent on annotation quantity and quality than model architecture. The main contributions of this paper are the classification of multiple PINTs simultaneously, and the improvement of PINTs classification for the German language.

4.1.2 Introduction

The inclusion of PINTs in synthetic speech can improve naturalness and intelligibility. For synthetic speech, pauses have been shown to improve digit recollection (Elmers et al., 2021a), whereas breath noises improve sentence recollection (Elmers et al., 2021b). The detection and modeling of breath groups can improve the quality of speech synthesis (Székely et al., 2019b, 2020). Previous work (Henter et al., 2016) has indicated the importance of quality training data for TTS applications. Most modern TTS systems are unable to generate PINTs with appropriate location, duration, and frequency, especially for spontaneous conversational situations. Similar to Székely et al. (2019b), an additional goal of this work is to incorporate this detection method into a future TTS pipeline, for generating appropriate PINTs for spontaneous synthesis. These TTS systems can then be incorporated further into robotics, call centers, digital agents, etc.

Often PINTs co-occur with one another in a variety of sequences. Condron et al. (2021) showed that training with more classes improved performance for non-verbal vocalizations (similar to PINTs) and laughter detection. The traditional approach has been to search for a single PINT, while collapsing all other PINTs to an 'other' class, or ignoring them altogether. Since these particles are not usually detected together, there is an absence of studies that incorporate state-of-the-art methods for detecting multiple PINTs simultaneously, especially for the German language. We expect that the classification of PINTs will benefit from simultaneous modeling, by training with multiple classes of PINTs, and have a positive outcome on synthesis quality for future research.

There are many applications for audio classification including medical, automatic speech recognition (ASR), and TTS. Previous classification research has distinguished between coughs and breath noises (Coppock et al., 2021), and detected respiratory disorders (Lei et al., 2014; Saraiva et al., 2020). Fukuda et al. (2018) found a reduction in error rate when using breath events as a delimiter for ASR, and Székely et al. (2020) found that annotating breath groups, and including breath noises, while omitting low probability breath events, created more fluent TTS.

Many methods have previously been used to detect PINTs: for silent segments (Braunschweiler & Chen, 2013; Singh et al., 2017; Garcia et al., 2018), breath noises (Székely et al., 2019b; Condron et al., 2021; Braunschweiler & Chen, 2013; Garcia et al., 2018), filler particles (Goto et al., 1999; Audhkhasi et al., 2009; Krikke & Truong, 2013; Reichel et al., 2019), and clicks (Condron et al., 2021; Garcia et al., 2018). PINTs classification has used a variety of methods, such as convolutional neural networks (CNN) (Saraiva et al., 2020), support vector machines (SVM) (Garcia et al., 2018), Gaussian mixture models (GMM) (Krikke & Truong, 2013), decision tree algorithms (Germesin et al., 2008), and template matching (Ruinskiy & Lavner, 2007; Lu et al., 2020).

In a pilot study conducted with a small English dataset, a neural network (NN) was used to perform a binary classification, predicting breath noises using mel-frequency cepstral coefficients (MFCCs) as input. Historically, MFCCs have performed well for audio classification. This approach appeared promising for the task of locating PINTs. Machine learning algorithms are extremely prevalent in current research. This paper will model PINTs using a NN, a CNN, and a recurrent neural network (RNN). The RNN is expected to outperform the other models since it is able to evaluate the temporal relationship between different PINTs.

4.1.3 Methods

Corpus

The Pool corpus (Jessen et al., 2005) consists of 100 male native speakers of German (age range 21–63 years old; mean age 39 years old). The present study considers the combination of the free technical setting with the spontaneous speech task, i.e. a picture description task. Similar to the board game Taboo, the speaker must describe a picture while not using any of the words listed beneath the picture.

This corpus has been annotated with information for different PINTs. There are 100 files in total (duration range 124–374 s; mean duration 223 s; total duration 6.2 hours). All signals are sampled at 16 kHz on a single channel. From these files, a total of 17,641 annotated PINTs were extracted (see Table 4.1). Additional classes were annotated like laughter, nasal filler particles (hm), glottal reflex, and other disfluencies like lengthening, truncation, and repair. However, their occurrences were too infrequent to include in the modeling.

Data Pre-Processing

The first step for pre-processing was to extract 13 MFCCs with a frame size of 93 ms, and a hop length of 23 ms, using the *Librosa* python package (McFee et al., 2015). Where the files differed in duration zero-padding was used in order to maintain the

Table 4.1: Overview of annotated PINTs. The minimum (min), maximum (max), mean, and standard deviation (sd) are measured in seconds. Total refers to the durational total and is measured in minutes. The proportion (prop) is the PINTs durational total divided by the total time of the corpus and is expressed as percentage out of 100%.

	count	**min**	**max**	**mean**	**sd**	**total**	**prop**
silent segment	10237	0.01	20.01	0.65	0.95	111.04	29.92
inhalation	2891	0.05	2.10	0.51	0.27	24.79	6.68
exhalation	1887	0.03	3.23	0.38	0.28	12.15	3.27
filler (uh)	1156	0.04	1.44	0.35	0.16	6.81	1.83
filler (uhm)	549	0.15	2.64	0.53	0.25	4.85	1.30
click	921	0.00	0.50	0.06	0.05	0.96	0.25

same size for modeling. The models were trained on the following nine classes: silent segments, inhalation, exhalation, two FPs (“uh” and “uhm”), clicks, task change (long stretches of silence while the interviewer changes tasks), zero-padding, and a final category for speech.

Model Architecture and Training

Models were implemented using *Keras* (Chollet et al., 2015). All models are compiled using a sparse categorical cross entropy loss function, a learning rate of 0.0001, the Adam optimizer, a batch size of 32, and for 40 epochs. A training/test split of 75/25 is used for all the models. Additionally, 20% of the training set is used for validation. Since there are 100 files, 60 files of material were randomly selected for training, 15 files of material were randomly selected for validation during model training, and 25 files of material were randomly selected and withheld for testing. Each model was trained using a different training/test split.

Neural Network

The NN model (see Fig. 4.1) incorporates a flattened input layer followed by two fully connected hidden layers, each with 64 neurons, a rectified linear unit (ReLU) activation function, and a 30% dropout for each layer. The output is a softmax layer to predict the output class. Training time is approximately 25 minutes on CPU.

Convolutional Neural Network

The CNN model (see Fig. 4.2) is comprised of two 1D convolutional layers. Each with 32 filters (size = 1, stride = 1), a ReLU activation function, followed by a 1D max pooling and batch normalization. The output is then flattened and fed into a

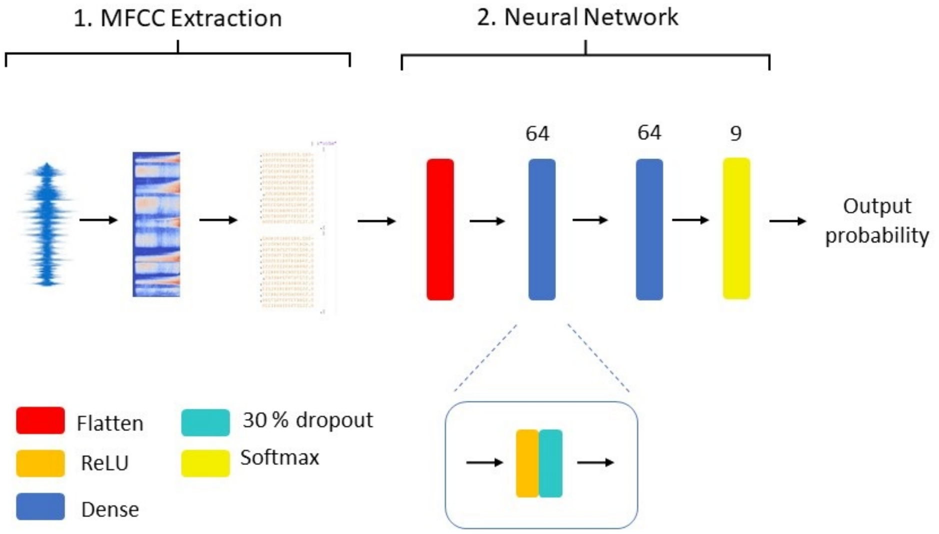


Figure 4.1: Architecture of NN.

dense layer with 64 neurons and a ReLU activation function, with a dropout of 30% applied to this layer. The output is a softmax layer for predicting the output class. Training time is approximately 35 minutes on CPU.

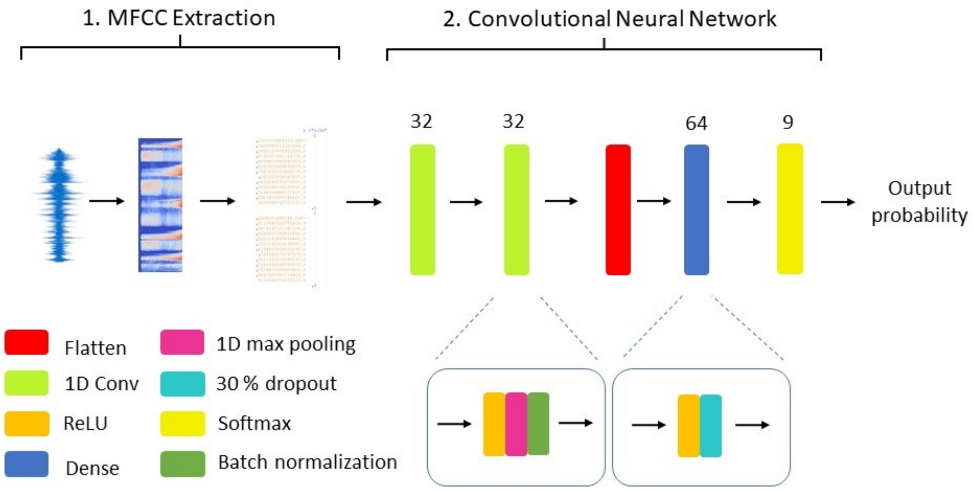


Figure 4.2: Architecture of CNN.

Recurrent Neural Network

The RNN model (see Fig. 4.3) consists of two fully connected long short-term memory (LSTM) layers each with 64 neurons. Next is a dense layer with 64 neurons, a ReLU activation function, and a 30% dropout. The output is a softmax layer which predicts the output class. Training time is approximately 70 minutes on CPU.

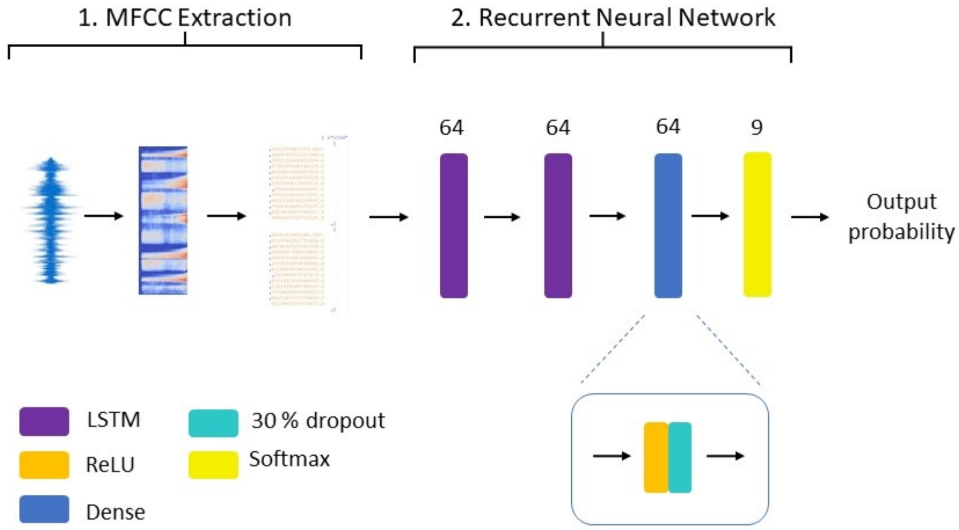


Figure 4.3: Architecture of RNN.

4.1.4 Results

Table 4.2 compares the accuracy, precision, recall, and F1 score for the three models. Both the CNN and the RNN performed slightly better than the NN in terms of accuracy and F1 score. The CNN and RNN performed similarly, except that the RNN performed better for precision. All models began with a relatively high accuracy and improved minimally throughout the remaining epochs. Overall, the scores for precision, recall, and F1 were lower than expected. Therefore, a confusion matrix was generated for each model (see Table 4.3, Table 4.4, and Table 4.5) to further investigate the classification of individual PINTs. All three models performed best when classifying silent segments, followed by inhalations and exhalations. For both inhalations and exhalations, they were most often confused for a silent segment in all models. Overall, the models performed well when separating inhalations from exhalations and vice versa. However, all models failed to classify FPs and clicks. Table 4.6 compares model performance for the individual PINTs. The CNN performed best when classifying silent segments, the NN performed best when classifying inhalations, and both the CNN and RNN performed equally well for classifying exhalations.

4.1.5 Discussion and Conclusion

This paper considered different machine learning architectures for classifying PINTs. Surprisingly, the NN, CNN, and RNN performed similarly, with some individual ad-

Table 4.2: Accuracy, Precision, Recall, and F1 Score for different models expressed as a percentage out of 100%.

	Accuracy	**Precision**	**Recall**	**F1 Score**
NN	85.6	53.5	41.6	40.5
CNN	86.1	53.2	41.9	41.8
RNN	86.1	69.0	42.1	41.7

Table 4.3: NN confusion matrix for test set. Rows correspond to annotated class and columns correspond to prediction.

	sil	**in**	**ex**	**uh**	**uhm**	**click**	**sum**
silent segment (sil)	64971	2743	789	-	-	-	68503
inhalation	4141	10372	58	-	-	-	14571
exhalation	3215	497	2188	-	-	-	5900
filler (uh)	60	3	34	-	-	-	97
filler (uhm)	68	4	33	-	-	-	105
click	209	85	6	-	-	1	301
sum	72664	13704	3108	-	-	1	89477

Table 4.4: CNN confusion matrix for test set. Rows correspond to annotated class and columns correspond to prediction.

	sil	**in**	**ex**	**uh**	**uhm**	**click**	**sum**
silent segment (sil)	66494	1375	754	-	-	1	68624
inhalation	5111	9351	100	-	-	-	14562
exhalation	3173	336	2532	-	-	-	6041
filler (uh)	53	2	27	-	-	-	82
filler (uhm)	80	5	20	-	11	-	116
click	181	73	11	-	-	-	265
sum	75092	11142	3444	-	11	1	89690

Table 4.5: RNN confusion matrix for test set. Rows correspond to annotated class and columns correspond to prediction.

	sil	**in**	**ex**	**uh**	**uhm**	**click**	**sum**
silent segment (sil)	64771	1813	811	-	-	-	67395
inhalation	4214	10098	113	-	-	-	14425
exhalation	2812	394	2308	-	-	-	5514
filler (uh)	38	2	13	-	-	-	53
filler (uhm)	50	2	17	-	3	-	72
click	165	74	8	-	-	3	250
sum	72050	12383	3270	-	3	3	87709

Table 4.6: Proportion correct for each model and class expressed as a percentage out of 100%.

	sil	**in**	**ex**	**uh**	**uhm**	**click**
NN	94.8	71.2	31.1	0.0	0.0	0.3
CNN	96.9	64.2	41.9	0.0	9.5	0.0
RNN	96.1	70.0	41.9	0.0	4.2	1.2

vantages in different cases. We had hypothesized that the RNN would perform better than the other two models since it is better able to capture temporal information. However, this was not the case. The models were able to easily identify silent segments and could classify inhalations fairly well, most likely due to them being the most frequently annotated classes. The models had middling success when attempting to detect exhalations. This is possibly due to the lower frequency of occurrence of exhalation annotations in the data. Inhalations and exhalations were sometimes misclassified as silent segments, possibly due to their frequent proximity.

All models were unable to accurately classify FPs and clicks. This finding is counter to the hypothesis that modeling multiple PINTs simultaneously would improve the classification accuracy of other PINTs. The models might have had difficulty classifying FPs because they were too similar to the speech category. The models struggled to properly classify clicks, which were often incorrectly classified as a silent segment. This could be in part due to the extremely short duration of clicks or a drawback of using only MFCCs as input.

The models were designed to encourage parity between them by having a similar number of layers, neurons for those layers, and the same hyperparameters. This decision was made to highlight the architectural differences of the models. During training time all three models started with a relatively high accuracy and only improved slightly during subsequent epochs. Since all the models performed similarly, we hypothesize that further improvements in accuracy could be gained by increasing the number of occurrences for the PINTs, especially the infrequent ones, showcasing the importance of quality annotations. Other possibilities include using techniques such as oversampling the less frequent PINTs, or undersampling the more frequent PINTs, to create a more balanced dataset. These techniques might increase the model's capability to correctly classify infrequent PINTs. In addition to MFCCs, other acoustic features should be investigated in order to improve classification. Since the inputs were MFCCs, the CNN model used 1D convolutional layers. Classification could possibly be further improved, by using spectrogram images instead of MFCCs for models using a CNN architecture. Future work could also investigate hybrid models that include the strengths of CNNs (i.e., feature extraction) and RNNs (i.e., temporal dependencies) to better detect and classify PINTs.

A primary goal for developing these classification models is to improve speech synthesis. Future work will implement a PINTs classification method as part of the training process for a TTS pipeline to create more natural, conversational speech synthesis.

4.2 Synthesis after a Couple PINTs: Investigating the Role of Pause-Internal Phonetic Particles in Speech Synthesis and Perception

4.2.1 Abstract

Pause-internal phonetic particles (PINTs), such as breath noises, tongue clicks and hesitations, play an important role in speech perception but are rarely modeled in speech synthesis. We developed two text-to-speech (TTS) systems: one with and one without PINTs labels in the training data. Both models produced fewer PINTs and had a lower total PINTs duration than natural speech. The labeled model generated more PINTs and longer total PINTs durations than the model without labels. In a listening experiment based on the labeled model we evaluated the influence of various PINTs combinations on the perception of speaker certainty. We tested a condition without PINTs material and three conditions that included PINTs. The condition without PINTs was perceived as significantly more certain than the PINTs conditions, suggesting that we can modify how certain TTS is perceived by including PINTs.

4.2.2 Introduction

PINTs are largely unconscious both in their production and perception and we are still working towards understanding the scale of their influence. PINTs, as speech planning tools or hedging mechanisms, are less salient than features such as focus, prominence, intonation or even "tone of voice". As conversation systems strive to be lifelike and realistic at all levels (Aylett et al., 2022), it is important to understand the functions of PINTs. If conversational systems begin using PINTs in their verbal interaction as a means of sounding more realistic and relatable, they risk producing PINTs that modify the perception of the message, like transmitting incorrect speaker certainty (Kirkland et al., 2022).

The goal of this study is to model PINTs based on a spontaneous speech corpus, and apply the resulting synthetic speech in a perceptual experiment. First, we present a technological contribution that incorporates PINTs from spontaneous speech into a TTS system. While synthesis of filled pauses and breath events have been the focus of other studies (Dall et al., 2016; Székely et al., 2019a, 2020), to the best of our knowledge, this is the first synthetic voice that is able to produce discourse clicks. Second, we demonstrate that a variety of PINTs patterns, generated with TTS, can be used as experimental material. This is a contribution to an emerging methodology that uses state-of-the-art neural TTS for stimuli creation, instead of manual manipulations of recorded speech samples (Kirkland et al., 2022). Specifically, we evaluated the effect of PINTs, on perceived certainty of the speaker, via a listening experiment.

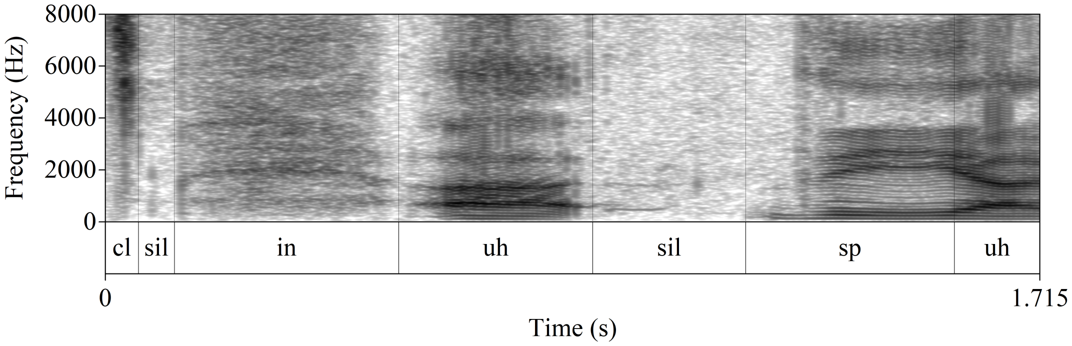


Figure 4.4: Example section from speaker. Annotations of PINTs: silence (sil), inhalation noise (in), exhalation noise (ex), filler particles (uh) and (um), tongue click (cl). Speech is annotated as “sp”.

4.2.3 Method

TTS Generation

Our training material is from Open Yale Courses (2007b), which is a project that provides free and open access to a number of introductory courses from Yale University. We selected lectures that included a high number of spontaneous speech phenomena. Next, we annotated a subset[1] of lectures totaling 3 hr 7 min for a single speaker with a diverse PINTs profile. The selected speaker’s PINTs material was approximately 40% of the total lecture time. An example annotation can be found in Figure 4.4.

Our training data incorporated transcripts taken from the Open Yale Courses website. We removed all punctuation in the original transcripts, as these are meant to improve the readability and do not correspond to acoustics. Next, we assigned PINTs to the available punctuation labels. For example, silence (,), inhalation (;), exhalation (.), tongue click (tk), filler particle (uh), and filler particle (um). The following is an example transcript with PINTs punctuation inserted: “; the metropolis which uproots people . , tk uh takes them away takes them out of ; traditional cultures , tk ;”. Numbers were typed alphabetically (e.g., nineteen twenty two), accented symbols (e.g., Leger vs. Léger) and hyphens (e.g., self consciously) were removed, and acronyms were written out (e.g., r i s).

The original annotations included an “other” category, which comprised a variety of phenomena such as laughter. The “other” labels from the annotations were not included in the training transcript because they comprised rare cases that were too infrequent to reliably model. We exclusively used punctuation and textual labels for PINTs, as opposed to introducing new symbols or phonemes. This ensures that our TTS system is capable of interpreting automatically generated input that is trained on text alone. In particular, this enables the fine-tuning of large language models on

[1] Lectures 1, 7, 13, and 24 from https://oyc.yale.edu/english/engl-310

TTS corpora, as demonstrated in Wang et al. (2022), to generate synthesis prompts that produce the distribution of PINTs in the training data. For example, inserting semi-colons in places where the speaker is likely to take a breath, or 'tk' tokens when a speaker is likely to use a tongue click.

The training data was segmented into breath groups following Székely et al. (2019b, 2020), which meant that audio snippets began and ended with an inhalation label. If the duration of the utterance was greater than 11 seconds, a constraint of Tacotron 2 (Shen et al., 2018), the audio was cut at a silence label instead. PINTs are often modeled beyond single sentences. However, due to the limitations of Tacotron2, we've only investigated single-sentence environments. All utterances were at least 4 seconds long. In total, we included 1224 breath group utterances, with 1128 in the training data and 96 held out for validation.

The TTS system was trained using a PyTorch implementation[2] of the sequence-to-sequence neural TTS engine Tacotron 2 (Shen et al., 2018). The models (with 28.2M parameters) were trained using transfer learning on a pre-trained model based on a large read speech corpus, LJSpeech (Ito & Johnson, 2017). This approach has been beneficial to TTS quality when training on a limited size spontaneous corpus. Specifically, in reducing the number of mispronunciations and increasing speed of convergence (Székely et al., 2019a). We trained two models on the data: ControlledPINT, where all transcribed PINTs are included with their own lexical token, and AutoPINT, where we removed the transcriptions of the PINTs. Phoneme-level input is used for training and synthesis and is obtained from the transcripts using the `g2p_en` package (Park & Kim, 2019). Both voices were trained for 70k iterations on top of the published read speech model, on 3 GPUs each, for 67 hours, with a batch size of 28. The speech signal is decoded from the model output using the neural vocoder HiFi-GAN (Kong et al., 2020).

To evaluate the PINT insertion of the ControlledPINT and the AutoPINT models, we compared their outputs to natural speech using five sentences that were excluded from the training data. For the ControlledPINT model, we designed the input to match the type and location of the PINTs in the natural sentence. The AutoPINT model used only the textual material. We synthesized multiple versions using each model and selected the versions with minimal distortions or errors, without regard to PINTs production. We avoided versions[3] that included metallic reverberations that would sometimes occur due to the recording conditions.

Perceptual Study

Using synthesized samples generated by the ControlledPINT model, we developed a perceptual experiment that uses generated audio to evaluate how PINTs influence

[2] https://github.com/NVIDIA/tacotron2

[3] Sample audio used for TTS comparison and perceptual experiment can be found at https://mikeyelmers.github.io/paper_interspeech23ttsdemo/

Table 4.7: Description of conditions used in perceptual study. The inserted material (punctuation labels) during generation is included.

condition	punctuation
PINTsless	*N/A*
long silence	, , ,
filler particle	, *um*
combinatory	, *um tk in*

certainty scores. In this study the participants listened to audio samples and evaluated how "certain" the speaker sounded of their opinion.

The textual material consisted of 10 sentences of similar syntactic structure, where the speaker describes their observations and opinions about artwork. For example, "The brush strokes in this painting contribute to a feeling of liveliness and energy". The semantic content of the utterances allowed for perceived hedging, indicating uncertainty. A Likert scale was used for evaluation with 1 representing "completely uncertain" and 7 representing "completely certain". Listeners heard a total of 40 audio stimuli, consisting of 10 different sentences synthesized in 4 different conditions (see Table 4.7). The "PINTsless" condition did not insert PINTs during synthesis. The "long silence" condition inserted a longer silence by including 3 silence symbols in a row. The "filler particle" condition inserted a silence and "um". And the "combinatory" condition inserted a silence, um, tongue click, and inhalation. Sentence final inhalations were removed, since the stimuli were evaluated in isolation. The tongue click in the "combinatory" version was surrounded by other PINTs because previous research has found that they co-occur alongside other PINTs in word-searching (Ogden, 2013; Moreno, 2019).

Our first hypothesis was that the PINTsless condition would be rated as more certain than the conditions that included PINTs. Our second hypothesis was that the combinatory condition would be rated as more certain than the filler particle condition. A FP may indicate that the speaker has encountered word search (e.g., lexical retrieval) problems and the following tongue click may signal that the word was found. The long silence condition was included as a distractor to prevent the participants from developing overly simple heuristics in their certainty ratings.

In an initial questionnaire, participants were asked about hearing impairment and age. All participants listened to the same set of stimuli, the order of which was randomized. The experiment required the use of headphones. Participants were asked to rate *'How certain does the speaker sound?'* on a 7-point Likert scale. The audio began automatically and the participants could click a "Play" button to hear the audio up to two more times before making their decision.

The perception study was created using a web-based experiment platform, Lab-

Table 4.8: Duration information for the different TTS models and natural speech for five sentences excluded from training. Both the total PINTs duration (PINTs dur) and the total audio duration (total dur) are measured in seconds. The proportion (prop) is measured out of 100%.

condition	PINTs dur	total dur	prop
natural	15.96	41.57	38.39
ControlledPINT	13.82	40.82	33.86
AutoPINT	7.95	36.11	22.00

vanced[4] (Finger et al., 2017), which presented the audio material and collected responses. We recruited participants using the crowd-sourcing platform Prolific[5] (2014). Fifty native English participants from the UK participated (mean age 40.7 years; age range 20–70 years, reflecting a diverse range of ages that represents a broad population). None of the participants self-reported hearing impairment. Participants were paid for their participation.

4.2.4 Results

Evaluation of TTS Model Performance

Using the five sentences that were excluded from the training data, we annotated three versions and measured the duration of their PINTs material (see Table 4.8). Our results[6] showed that the natural condition had the longest duration of PINTs material, which closely matched the overall PINTs profile proportion of the speaker (40%). The ControlledPINT model produced the second longest durations and second largest proportion, while the AutoPINT version produced the shortest durations and smallest proportion. These findings were expected, but it was noteworthy that the Controlled-PINT version closely resembled the natural version, and that the AutoPINT version could generate PINTs durations and proportion that were half of natural speech without any explicit labels. This is in line with the findings of Székely et al. (2019a), where filled pauses were automatically synthesized with a similar method.

We also looked at count information for the individual PINTs grouped by condition (see Table 4.9). The natural condition has the highest count values, with many more silences, especially edge silences that are adjacent to other PINTs, than material generated by either of the TTS systems. The ControlledPINT model produces more of the filler particle "uh" than the AutoPINT condition, but both systems produced the same number of "um" filler particles. The ControlledPINT system sometimes

[4] Accessed via https://www.labvanced.com/ on Feb. 24, 2023.

[5] Accessed via https://www.prolific.co/ on Feb. 24, 2023.

[6] All data and code for the results can be accessed at https://github.com/MikeyElmers/paper_interspeech23

Table 4.9: Count information for the different TTS models and natural speech: silence (sil), inhalation (in), exhalation (ex), filler particles (uh) and (um), tongue click (cl), and other (o).

condition	sil	in	ex	uh	um	cl	o
natural	43	23	2	10	–	2	8
ControlledPINT	14	14	1	17	1	–	–
AutoPINT	10	13	1	4	1	–	–

produce multiple PINTs from a single label, rendering more "uh" PINTs than was present in the natural speech. Only the natural material had tongue clicks or other labels.

Both the ControlledPINT and AutoPINT systems sometimes generate exhalations without a label. These exhalations were often near other PINTs in the data and this close association might be the cause of their unlabeled inclusion. Further evidence to support this theory comes from the ControlledPINT system sometimes producing a sequence of PINTs from just one or two labels in the input. Overall, the system is able to generate PINTs well, mirroring the PINTs pattern of the speaker. Occasionally, in cases with 5 or more PINTs in a row, the system struggles to perfectly recreate the PINTs sequence.

Our observations also revealed that tongue clicks were only realized by Tacotron2 when adjacent to silences or breath events. This is likely due to the fact that tongue clicks were one of the rarer PINTs in the training data and were almost always adjacent to other PINTs. Without an inhalation or silence in the prompt, the synthesizer would attempt to pronounce the tongue click symbol (tk) phonetically. The quality and loudness of the synthesized audio was also variable, likely due to differences in recording conditions across lectures. Originally, we expected the models to be quite probabilistic in their PINTs generation, however, they were more consistent than expected. Sometimes versions differed in their PINTs content but more often the differences were due to prosody and pronunciation variations.

Evaluation of the Perceptual Study

We created material, generated by the ControlledPINT system, for a perceptual experiment to evaluate the certainty of sentences in four conditions. The results for the perceptual study are in Table 4.10. Participants used the full scale in all conditions. We incorporated three measures of central tendency: mean, median, and mode. Each of these measurements highlights a different aspect of the data. For example, the mode for the PINTless condition was 7, indicating that the most common value was the highest possible rating.

The results confirmed our initial hypothesis that the PINTsless version would

Table 4.10: Descriptive statistics for the different conditions.

conditon	mean	median	mode	sd
PINTsless	5.90	6	7	1.10
long silence	4.31	4	4	1.30
filler particle	3.72	4	4	1.24
combinatory	3.51	3	4	1.27

Table 4.11: ANOVA comparison of base model and model with condition (cond) as a predictor. Includes number of parameters (par), AIC, log-likelihood (logLik), likelihood ratio test statistic (LR), degrees of freedom (df), and p-value (p).

	par	AIC	logLik	LR	df	p
base	8	5718.5	−2851.2	–	–	–
cond	11	5637.2	−2807.6	87.24	3	< 0.001

be rated more certain than the conditions with PINTs. The mean, median, and mode all clearly indicate that the PINTsless version sounded most certain. We also hypothesized that the combinatory version would be rated as slightly more certain than the filler particle condition. However, the data did not support this. All three PINTs conditions had similar certainty ratings but the long silence condition had the highest mean of the three PINTs conditions. The long silence condition also had more certainty scores in the 5-7 range than the other two PINTs conditions. The certainty scores for both the filler particle and combinatory conditions are similar but the filler particle condition has marginally higher certainty scores.

Statistical modeling was conducted with cumulative link mixed models (clmm) from the ordinal (Christensen, 2022) (Version 2022.11-16) package in R (R Core Team, 2021) (Version 4.1.2). A post-hoc analysis was conducted using emmeans (Lenth, 2023) (Version 1.8.4-1).

We compared a base clmm model, $clmm(certain \sim (1 \mid id) + (1 \mid stimuli)$, and a condition model, $clmm(certain \sim condition + (1 \mid id) + (1 \mid stimuli)$. The condition model predicts the certainty score with a single predictor, condition, as a fixed effect. For random effects, both subject id and stimuli with intercepts was included. An `anova()` was used to compare the two models. Table 4.11 shows that the model with condition as a predictor provides a significantly better fit than the base model as determined by both the Akaike information criterion (AIC) (Akaike, 1973) and log-likelihood.

A post-hoc analysis for pairwise comparisons was conducted (see Table 4.12). The PINTsless condition was significantly different ($p < 0.001$) from all PINTs conditions. The long silence condition was significantly different from both the filler particle

Table 4.12: Post-hoc pairwise comparison using Tukey method for comparing a family of 4 estimates.

contrast	est	SE	df	z	p
$fluent - sil$	3.320	0.312	Inf	10.624	< 0.001
$fluent - tc$	4.852	0.321	Inf	15.130	< 0.001
$fluent - um$	4.409	0.318	Inf	13.859	< 0.001
$sil - tc$	1.532	0.302	Inf	5.077	< 0.001
$sil - um$	1.089	0.301	Inf	3.621	< 0.01
$tc - um$	-0.443	0.300	Inf	-1.478	0.451

Table 4.13: Model information.

	est	SE	z	p
sil	-3.32	0.31	-10.62	< 0.001
click	-4.85	0.32	-15.13	< 0.001
um	-4.41	0.32	-13.86	< 0.001

condition ($p < 0.01$) and the combinatory condition ($p < 0.001$). However, the filler particle and combinatory conditions were *not* significantly different ($p = 0.451$).

Table 4.13 shows that the three PINTs condition (sil, click, and um) are all significantly different from the reference level which is the fluent condition. Table 4.14 shows the 95% confidence intervals for the different PINTs conditions which have overlap.

4.2.5 Discussion

In our listening experiment, we expected the combinatory condition to indicate higher certainty scores than the other PINTs conditions, but this was not the case. Surprisingly, the long silence condition received slightly higher certainty scores than the other two PINTs conditions. One possible explanation is that the long silence condition might be less obtrusive than the filler particle or combinatory conditions. However, the long silence condition still disrupts the flow of speech more than the PINTsless

Table 4.14: 95% Confidence Intervals.

condition	2.5%	97.5%
sil	-3.93	-2.71
click	-5.48	-4.22
um	-5.03	-3.79

condition, thereby reducing the listener's certainty. All PINTs were evaluated equally even though each PINT has different realizations that can influence certainty. Additionally, evaluating the effects of dialect, age, and gender for the interpretation of PINTs was outside the scope of this experiment.

Tongue clicks exhibit a number of functions such as: introducing a new sequence or topic, word search, maintaining a turn, backchanneling, stance marking, and repair (Ogden, 2013; Zellers, 2022). The acoustic realizations of these tongue clicks are highly variable (Trouvain & Malisz, 2016), which means that the tongue clicks the TTS engine rendered might behave differently from our intended function. The fact that tongue clicks did not improve certainty by signaling a successful word search affirms that the production and perception of different PINTs patterns might require more elaborate experimental design, such as in-context perceptual evaluations. Future research could provide insights for audio enhancement tools, to reveal which tongue clicks can be removed from the recording and which are necessary for retaining the speaker's original intent.

We created two different TTS systems that were able to produce PINTs. The annotations for the TTS corpus were made manually, and therefore a time-consuming process. One limitation of the manual annotations was that we were only able to evaluate a single speaker. Automatic detection of PINTs is a challenging task, especially since some particles are less common than others (Elmers, 2022). Improvements could be made by including more data and more consistent audio quality. This study modeled PINTs in single-sentence environments. Future work should explore multi-sentence environments, which are more representative of the way PINTs occur in natural speech. The experiment in our paper is one possible example of how generative modeling can be used to create materials and test hypotheses, in this case improving our understanding of the functional properties of PINTs. Using generative modeling to distill knowledge is not going to replace the need for corpus-based research, but it is becoming a useful and necessary addition.

4.2.6 Conclusion

We developed an annotation scheme that uses plain text punctuation symbols to describe a speaker's PINTs pattern, which focused on consistency for successful generative modeling. Using these annotations, we trained two synthetic voices: ControlledPINT and AutoPINT. ControlledPINT used overt PINTs labels in the training material. AutoPINT did not include any PINTs labels and relies on the probabilistic rendering of Tacotron2 to insert them automatically. We conducted a quantitative and qualitative analysis of these two models. The novelty of our models is that they are the first to produce tongue clicks. Using the output of the ControlledPINT model, we conducted a perceptual experiment to evaluate how certain a synthetic speaker sounds in 4 different conditions. Importantly, we have shown that by incorporating natural phenomena (e.g., clicks), we are able to create manipulated experimental material. We hope that this line of research will contribute towards a deeper understanding of these complex and latent speech phenomena. Additionally, by including controllable PINTs material in TTS voices, we can equip conversational systems to utilize PINTs, to better manage perception in social communication.

Chapter 5

PINTs Recall in Lecture Setting

The experiments contained in this chapter investigated the usage and perceptual effects of PINTs material in educational settings, specifically in lecture environments. Experiment 1 (chapter 5.1) analyzed the differences in PINTs between lecture material from Yale University and the TOEFL iBT listening practice section. We found that PINTs have a strong presence in both of the aforementioned lecture environments. This led to evaluating the recall benefits of PINTs, using lecture materials, in the second (chapter 5.2) and third (chapter 5.3) experiments. Experiment 2 and 3 used the same experimental methodology, with the major difference being that experiment 2 used natural materials, while experiment 3 used synthetic materials.

5.1 Comparing PINTs in University Lectures

5.1.1 Abstract

This study compared the PINTs usage of five different lecturers recorded at Yale University to the lecture listening sections of the TOEFL iBT English-language test. We annotated 5 hours of material from Yale lectures and 15 minutes of material from the TOEFL iBT lecture listening section. Additionally, we evaluated intra-speaker PINTs variation for the Yale lectures since the recordings spanned a three month semester. The following PINTs were annotated: silences, inhalation breath noises, exhalation breath noises, the filler particles “uh” and “um”, tongue clicks, and an “other” category. Results showed that PINTs comprised approximately 30% of the total time for Yale lectures, and 20% for the TOEFL iBT lecture listening section. Each Yale lecturer showcased a different PINTs pattern with respect to type and number of PINTs, but remained individually consistent with their PINTs type and frequency during the semester. Our findings showcase a need for additional research, especially on how PINTs influence recall, since PINTs can inhabit approximately 1/3 of the total lecture time.

5.1.2 Introduction

PINTs provide a great deal of information, and despite their relatively short duration, they exert a large influence upon style (Trouvain & Barry, 2000), perception (Bosker et al., 2014), and memory (Corley et al., 2007). University lectures showcase unique, style-specific features. For example, lectures are semi-prepared and somewhat-rehearsed while simultaneously spontaneous and monologic. This is in contrast to sentences recorded in a laboratory setting with a neutral style. Regarding lectures, Kjellmer (2003, p.190) stated, "a lecture that is read aloud from the written page is often difficult to take in when its delivery lacks the verbal guides and signposts that we more or less subconsciously expect to find in speech; as listeners we are in danger of missing the point of the argument."

In this study we annotated lectures and provide a descriptive statistical analysis of the PINTs information. Specifically, we'll look at PINTs count, type, duration, and correlation. Exploring where, what kind, and how often PINTs are present in lectures is an important first-step towards investigating the influence of PINTs on memory in a university lecture setting. Therefore, we've opted to also annotate and compare the Yale lecture annotations to annotations from the TOEFL iBT English-language lecture listening section. The study is interested in the follow:

- How often do lecturers use PINTs?
- Is PINTs usage unique to each lecturer?
- Does PINTs usage change for each lecturer during the three month semester?
- How does PINTs usage compare to an English-language lecture listening test?

5.1.3 Method

Lecture material was collected from Open Yale Courses (2007b), which contains free and open access courses provided by Yale University. We selected five lecturers (Hammer, 2007a; Wargo, 2010; Merriman, 2008; Wrightson, 2009; Echevarría, 2009) that used many spontaneous speech phenomena. We then annotated a subset of lectures totalling approximately 5 hours (i.e., 1 hour / lecturer). Since the lectures were recorded over a three month semester, 1/3 of the annotations were from the first lecture, 1/3 of annotations were from the half-way point of the semester, and 1/3 of annotations were from the final session of the semester. With multiple lecturers and data from three different time points, we were able to compare both inter- and intra-lecturer PINTs usage. The audio was annotated with the following PINTs labels: silence, inhalation breath noises, exhalation breath noises, the filler particles "uh" and "um", tongue clicks, and an "other" category (e.g., coughing, swallowing, etc.). No minimum threshold was enforced for any of the PINTs.

Table 5.1: Count information for the Yale and TOEFL lecture annotations: silence (sil), inhalation (in), exhalation (ex), filler particles (uh) and (um), tongue click (cl), and other (o).

	sil	**in**	**ex**	**uh**	**um**	**cl**	**o**
Yale	11559	5234	800	3218	277	625	471
TOEFL	473	192	12	58	24	20	1

The TOEFL iBT (Test of English as a Foreign Language, Internet-based Test) is a standardized test that evaluates the English-language proficiency for non-native speakers of English, and is a popular admission test for university entrance. We chose to compare the TOEFL iBT to the Yale lectures since the test has a focus on higher education preparation. In this study, we collected material from the TOEFL[1] iBT listening practice test. We annotated the same PINTs information for the TOEFL lectures as for the Yale lectures. The amount of material available for annotation for the TOEFL lecture listening practice was approximately 15 minutes, significantly less than what was annotated for the Yale lectures.

5.1.4 Results

R (R Core Team, 2021) (Version 4.1.2) was used to perform both descriptive and inferential statistics on the data[2]. Visualizations were generated using ggplot2 (Wickham, 2016) (Version 3.4.2).

We found 22,184 PINTs in the 5 hours of annotated Yale lecture material, compared to 780 PINTs found in the 15 minutes of annotated TOEFL lecture material. Table 5.1 shows count information for the individuals PINTs for both the Yale and TOEFL lectures. Both the Yale and TOEFL annotations show that a majority of annotations are made up of silences, inhalation, and the filler particle “uh”.

Duration

Duration information for each PINT type are provided for the Yale lecture data (see Table 5.2) and for the TOEFL lecture data (see Table 5.3). Individual PINTs duration information for the Yale lecture data can also be seen in Figure 5.1. The silence PINT type has a large standard deviation for both data collections, since silences inhabit a variety of durations.

[1] The TOEFL lecture listening practice was downloaded from https://www.ets.org/toefl/test-takers/ibt/prepare/practice-tests.html from the link titled “Download TOEFL iBT Listening Practice Sets with audio tracks (zip)” under the “TOEFL iBT Practice Sets” section.

[2] This project’s data and scripts can be found at https://github.com/MikeyElmers/paper_pp22.

Table 5.2: Duration information for each PINT type from Yale lecture annotations. The mean and standard deviation are measured in milliseconds. The total duration (total dur) for each PINT is measured in seconds. The proportion (prop) is measured out of 100%.

PINTs	**mean**	**sd**	**total dur**	**prop**
sil	197	319	2279	12.6
in	335	164	1754	9.7
ex	213	155	171	0.9
uh	262	124	842	4.7
um	410	129	114	0.6
cl	62	33	39	0.2
o	209	205	99	0.5

Table 5.3: Duration information for each PINT type from TOEFL lecture annotations. The mean and standard deviation are measured in milliseconds. The total duration (total dur) for each PINT is measured in seconds. The proportion (prop) is measured out of 100%.

PINTs	**mean**	**sd**	**total dur**	**prop**
sil	165	208	78	8.2
in	399	170	77	8.1
ex	252	138	3	0.3
uh	310	99	18	1.9
um	402	80	10	1.0
cl	34	26	1	0.1
o	143	–	.143	0.0

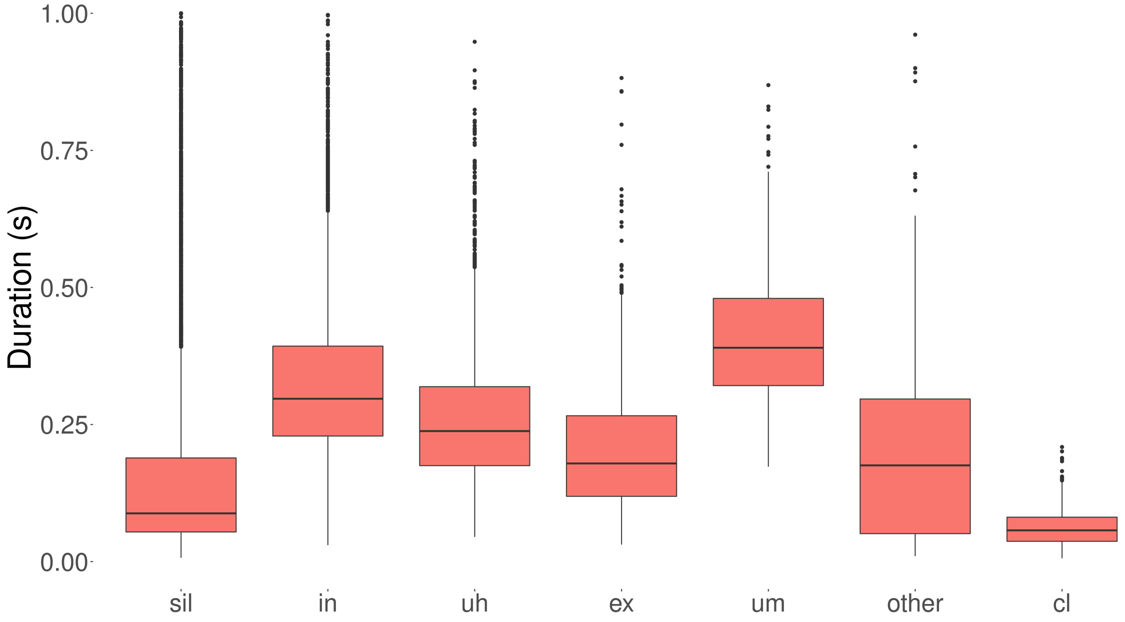


Figure 5.1: Box plot for individual PINTs duration.

Proportion

PINTs comprised approximately 30% of the total time for the Yale lectures (see Figure 5.2), and approximately 20% of the total time for the TOEFL lectures. Figure 5.3 shows a bar plot representation of the individual PINTs proportions for the Yale lecture data set. The total PINTs proportion is mostly compromised of silences, inhalations, and the filler particle “uh” for both data sets. The proportions for each PINTs type is usually higher in the Yale lecture data.

Inter-Lecturer Comparison

Each lecturer displayed a unique PINTs usage for the Yale lecture annotations. Table 5.4 shows duration differences for each of the five lecturers. The lowest PINTs proportion was approximately 20% for lecturer 3, and the highest PINTs proportion was approximately 40% for lecturer 1. This result indicates large differences between total PINTs duration for each lecturer.

Intra-Lecturer Comparison

We evaluated intra-lecturer PINTs usage (see Figure 5.4). Different speakers prefer different PINTs. For example, lecturer 1 uses many more “uh” compared to lecturer 4. Each lecturer tended to be consistent with their PINTs usage throughout the semester.

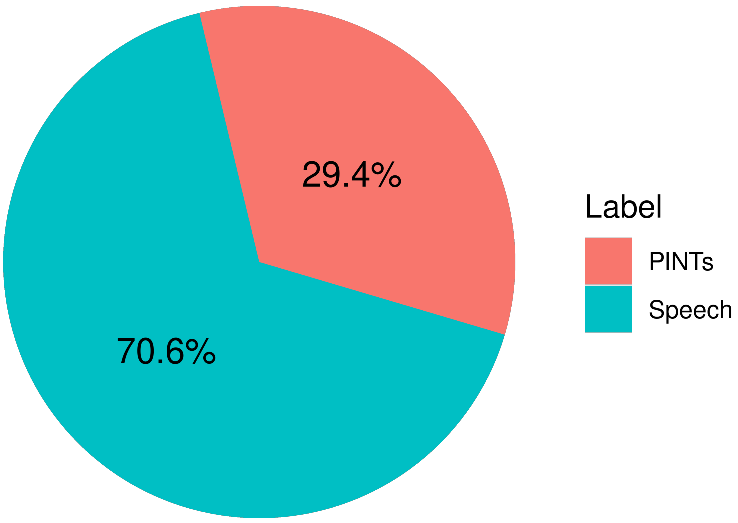


Figure 5.2: Pie chart for proportion of PINTs to speech.

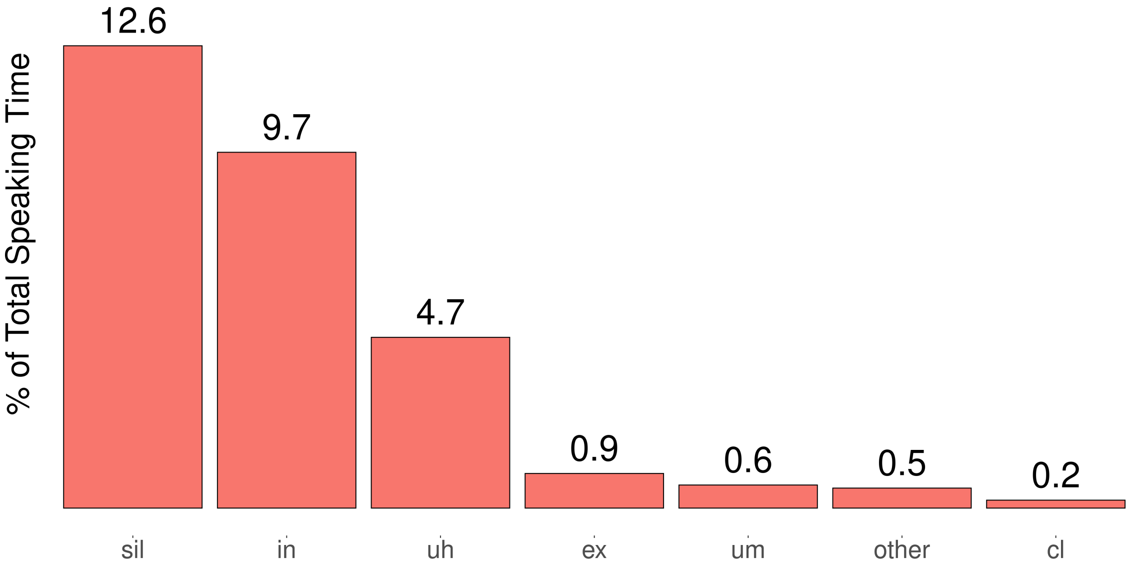


Figure 5.3: Bar plot for individual PINTs proportion out of total speaking time.

Table 5.4: Duration information for each of the five lecturers from the Yale lecture annotations. Both the total PINTs duration (PINTs dur) and the total audio duration (total dur) are measured in seconds. The proportion (prop) is measured out of 100%.

lecturer	PINTs dur	total dur	prop
1	1457	3604	40.4
2	937	3602	26
3	787	3609	21.8
4	990	3603	27.5
5	1126	3604	31.3

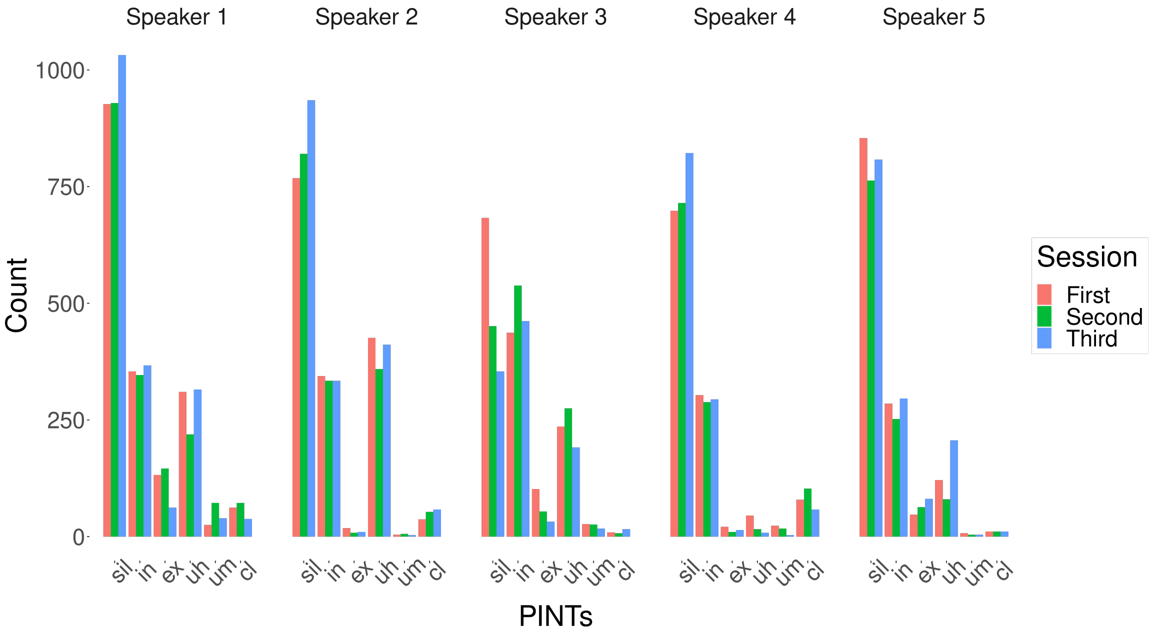


Figure 5.4: Bar plots for between- and within-speaker variability for each speaker and session.

Table 5.5: Comparison of speech rate and PINTs rate for each Yale lecturer.

lecturer	speech rate	PINTs rate
1	2.47	1.56
2	3.20	1.39
3	3.42	1.12
4	2.86	0.99
5	2.67	1.09

PINTs Rate

Table 5.5 compares the speech rate (i.e., number of syllables / total time), measured in the number of syllable per second, to the PINTs rate (i.e., number of PINTs / total), which is measured in the number of PINTs per second. Speakers varied in their PINTs frequency, with some having a PINTs rate as low as 0.99 PINTs/sec, and others with 1.56 PINTs/sec. The correlation analysis between speech rate and PINTs rate did not reveal a statistically significant relationship (t(3) = -0.55, $p > 0.05$, r = -.30).

5.1.5 Discussion

Since teaching time is valuable, we expected to find few, if any, PINTs during the Yale lectures. Instead we found many particles that accounted for more than 30% of the total time. This finding is in stark contrast with current speech synthesis techniques which include silences but omit the other PINTs. Meaning, current synthesis techniques ignore about 30% of material, at least for this speech genre. Additionally, these results are evidence against the belief that silence dominates pauses since approximately 2/3 of the entire pause duration consists of particles that are not silent. This work can function as a baseline for achieving a more natural usage of PINTs for lectures in speech synthesis.

While the PINTs usage was lower in the TOEFL lecture data, there numbers were still comparable to the low end of lectures in the Yale data. Again, we did not expect to find PINTs included in the TOEFL listening test, especially since the audio seems appears curated. It is important to investigate whether these particles have an influence on the recollection of lecture material, and if the effects are consistent for both native and non-native speakers.

5.1.6 Conclusion

The present study compared the PINTs usages for annotation from both Yale and TOEFL lectures. Both the Yale and TOEFL annotations showed a majority of counts comprised silences, inhalations, and the filler particle “uh”. We found that PINTs comprised approximately 30% of the total time for the Yale data, and approximately 20% of the total time for the TOEFL data. In most cases the PINTs proportions were higher in the Yale data. Results indicated unique PINTs usage in terms of type, total duration, and frequency for the five lecturers annotated in the Yale data. Despite the differences in the type, amount, and duration between lecturers, PINTs usage remains relatively consistent over the semester for each lecturer. The strong presence that PINTs can occupy in both testing and real-world lectures indicates a need continuing to investigate their influences.

5.2 Pause Particles Influencing Recollection in Lectures

5.2.1 Abstract

This study investigated the influence of pause-internal phonetic particles (PINTs) on recall for native and non-native listeners of English. Participants were 45 monolingual English and 45 L1 German listeners who heard segments from university lectures, in English, and answered content-based questions. Three versions of lecture stimuli were created: an unmanipulated original version, a "silence" version, and a "no PINTs" version where all PINTs were removed including silences. In the original and "silence" versions, half of the key information was preceded by PINTs material. The results indicated that material preceded by PINTs was less likely to be recalled. Additionally, the participant's first language was not significant for understanding the speaker. However, English listeners tended to score higher during the "no PINTs" condition, while German listeners tended to score higher during the original condition. This study was unable to replicate the recall benefit of PINTs found in single sentence laboratory setting experiments.

5.2.2 Introduction

Pause particles can exhibit an influence on recollection. For example, Fraundorf & Watson (2011) found that the recollection of story plot points was improved when including FPs. In word recognition studies, Corley et al. (2007) found that disfluencies improved the recollection of the following word, while MacGregor et al. (2010) found that silent pauses improved the recollection of the following word. Importantly, MacGregor et al. (2010) claims that a feature of disfluencies is that they provide additional time. Watanabe et al. (2005) found that native and non-native listeners exhibited shorter response times for complex phrases that were preceded by FPs or silence compared to a no pause condition. Overall, these studies show that PINTs can affect recollection in laboratory settings. However, these studies do not utilize material from a real-world setting and focus on smaller contexts (i.e., words or sentences). This study expected to find a PINTs benefit for recollection in university lecture segments, similar to the previously mentioned smaller contexts.

Similar to Wagner et al. (2015), this study does not advocate for 'lab speech' or 'natural speech', rather the goal is to improve awareness around the types of data and methods used. This study explored the influence of PINTs on memory, using real-world data, rather than in a laboratory setting and with material larger than a single sentence. Another main goal was to evaluate the effect of PINTs on both native speakers (NSs) and non-native speakers (NNSs). We opted to examine English monolingual listeners and L1 German listeners due to the English language stimuli used in the study.

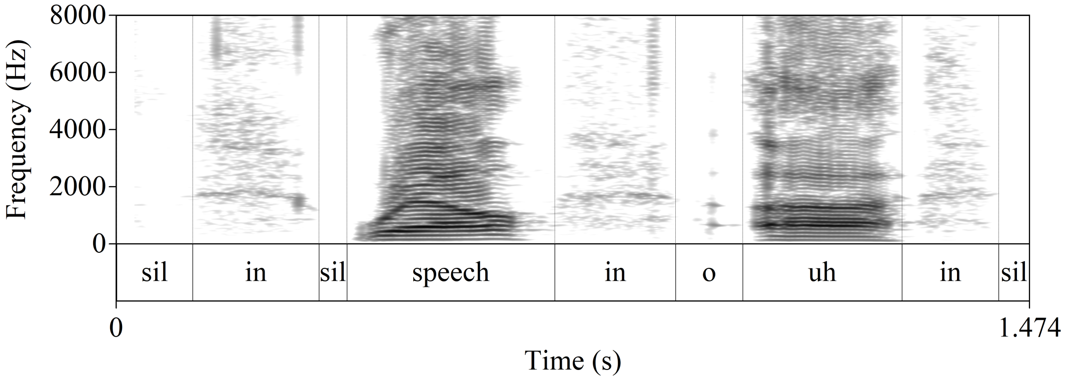


Figure 5.5: Example section from speaker. Annotations of PINTs: silence (sil), inhalation noise (in), filler particle (uh), and other (o).

5.2.3 Method

Lectures were collected from Open Yale Courses (2007b) which contains free and open access courses from Yale University. English-language lectures were chosen based on the speaker's PINTs profile. After selecting a specific speaker, annotations were made for a subset of their lectures. The chosen speaker displayed a relatively high number of PINTs during his lectures, with upwards of 40% of his total time incorporating PINTs material. Fig. 5.5 shows an example for this speaker.

Participants

This study used a web-based experiment created with Labvanced[3] (Finger et al., 2017) to present the audio stimuli to participants, and to collect their answers and questionnaire information. Participants were recruited using the crowd-sourcing platform Prolific[4] (2014) and consisted of 45 monolingual English participants (mean age 38 years; age range 21–62 years) and 45 L1 German participants (mean age 35 years; age range 21–72 years) who were paid for their participation. One monolingual English participant reported hearing impairment and was not included in the results.

Stimuli

Stimuli consisted of four three-minute sections extracted from full length lectures. Each audio segment was followed by two multiple-choice content-based questions, with one question preceded by PINTs material and the other not. The study was balanced so that the key material was equally preceded, or *not* preceded, by PINTs. However, neither question was preceded by PINTs material in the "no PINTs" condition, since all PINTs material was removed. An example question was: "According

[3] Accessed via https://www.labvanced.com/ on Dec. 06, 2022.

[4] Accessed via https://www.prolific.co/ on Dec. 06, 2022.

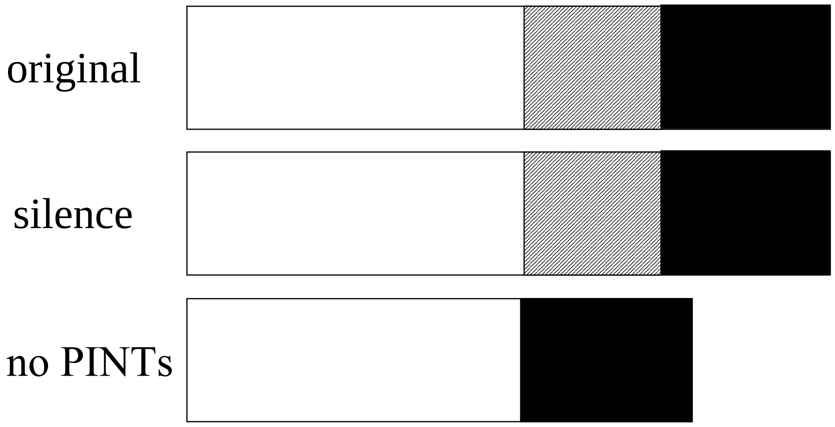


Figure 5.6: Schematic of the duration for the three conditions showing speech (white), PINTs (grey), and speech material that contained the key information (black).

to Paul Fussell, what is the essential trope or rhetorical figure of World War One poetry?" The possible answers were: a) hyperbole, b) metaphor, c) oxymoron, d) irony. The participants did not need to know what these concepts meant, or any other encyclopedic or background knowledge. Instead, they needed to answer based on the content as presented by the lecturer.

The different conditions were created using a Praat (Boersma & Weenink, 2022) script that would remove or replace the PINTs material. In the "silence" condition, non-silence PINTs were replaced with a silence taken from the audio and matched to the duration of the cut material. Therefore, the "silence" condition maintained the same duration as the original audio. The original and "silence" conditions provided the same amount of processing time, while the "no PINTs" condition provided less processing time (see Fig. 5.6). The "no PINTs" condition did not include any acoustic pause whatsoever. Participants only heard one of the three conditions, i.e., one third of participants heard four original clips, one third of participants heard four "silence" clips, and one third of participants heard four "no PINTs" clips. Each of the conditions included the same textual material, however, the order of the four audio clips was randomized to prevent ordering effects.

Procedure

Participants were informed that they would hear four audio clips, each approximately three minutes, and answer content-based questions immediately following each clip.

Table 5.6: Mean score, median score, standard deviation score, and count information for the different conditions and L1s.

condition	L1	mean	median	sd	n
noPINTs	*EN*	6.26	7	1.83	15
silence	*EN*	6.07	6	1.44	13
original	*DE*	6.00	6	1.07	15
original	*EN*	5.88	6	1.92	16
noPINTs	*DE*	5.87	6	1.41	15
silence	*DE*	5.67	5	1.76	15

Participants were instructed to use headphones and test their audio volume before starting. They were told to not take notes. They were also told that the recordings were from a non-ideal microphone and included some background noise. This was in order to draw their attention away from some of the minor artefacts that occurred from the audio manipulation in the "silence" and "no PINTs" conditions. Participants were told that they would receive a score at the end of the test as an additional incentive to perform well. While listening to the audio, participants saw "Listen closely!" on their screen. They heard each audio clip only once.

After completing the listening section, participants answered a questionnaire that included: age, hearing impairment, L1, self-assessed English skills (for the German listeners), a test score if possible (for the German listeners), highest completed education (high school, university, or other), level of interest in the audio contents (1: very uninterested to 5: very interested), how easy the speaker was to follow and understand (1: very difficult to 5: very easy), and how prepared they found the speaker (1: very unprepared to 5: very prepared). Total completion time was between 15-20 minutes.

5.2.4 Results

Participants were scored based on how many questions they answered correctly with a maximum score of 8 (1 point per correct answer). The monolingual English participants scored higher than the L1 German participants in all conditions except in the original condition, however, the monolingual English speakers usually had a higher variance (see Table 5.6). Monolingual English participants scored highest during the "no PINTs" condition while the L1 German participants scored highest during the original audio condition.

The data[5] was pre-processed using the dplyr (Wickham et al., 2023a) (Version 1.1.1), stringr (Wickham, 2022) (Version 1.5.0), and tidyr (Wickham et al., 2023b)

[5] This project's data and scripts can be found at `https://github.com/MikeyElmers/paper_icphs23`.

Table 5.7: Mean and standard deviation score based on whether the answer was immediately preceded by PINTs material.

preceding PINTs	mean	sd
no	0.81	0.39
yes	0.66	0.47

(Version 1.3.0) packages. Homogeneity of variance was evaluated using Levene's test from the car (Fox & Weisberg, 2019) (Version 3.1.2) package. Statistical models were analyzed with linear regression and binomial generalized linear mixed models (binomial GLMMs) with the lme4 (Bates et al., 2015) (Version 1.1.31) and lmerTest (Kuznetsova et al., 2017) (Version 3.1.3) packages in R (R Core Team, 2021) (Version 4.0.4). Models were compared with the Akaike information criterion (AIC) (Akaike, 1973) to calculate unexplained variance. The best fit model was selected as the model with the lowest AIC.

Preceding PINTs

This study investigated the effect of PINTs material immediately before key information on participant score. These models did not include the "no PINTs" audio condition since all PINTs material was removed. Scores are out of 1 rather than 8 since the evaluation is done on a by-question basis rather than a subject's collective score. Table 5.7 shows that when key information was preceded by PINTs, the result was an overall lower score. The data showed violations for normality, as indicated by the Shapiro-Wilk test, and homogeneity of variances, as indicated by Levene's test. Therefore, the non-parametric Wilcoxon rank sum test was used. Results indicated a significant difference between the preceding PINTs conditions (*W = 32096, p < 0.001*). Participants performed significantly better when critical information was *not* preceded by PINTs information.

Binomial GLMMs were used to evaluate which variables influenced score. The model with the best fit was: $glmer(score \sim precede + (1 \mid id), family = binomial)$. This model predicts score based on the answer being preceded by PINTs information as a fixed effect, and subject with intercept as a random effect. This model performed better than models that incorporated L1, condition, or the questionnaire variables. The analysis revealed a main effect for preceding PINTs (*Estimate = -0.88, SE = 0.23, z = -3.87, p < 0.001*). This main effect indicates that the presence of PINTs material before the answer lowered participants' score.

Ease

Participant's reported how easy it was to follow and understand the speaker (1: very difficult to 5: very easy). Overall, the mean ease was 2.82. The condition that re-

Table 5.8: Summary information of linear model with ease as predictor.

	est	**SE**	**t**	**p**
(*Intercept*)	4.84	0.42	11.60	< 0.001
*ease*2	1.15	0.51	2.28	< 0.05
*ease*3	0.95	0.54	1.77	0.08
*ease*4	1.54	0.55	2.81	< 0.01
*ease*5	1.95	0.64	3.09	< 0.01

Table 5.9: Pearson correlations between total score and questionnaire responses for all participants.

	age	**ease**	**interest**	**prep**
total score	0.09	0.30	0.10	0.28

moved all the pause material was considered the easiest to follow when averaging over all participants. However, an analysis of variance (ANOVA) showed no significant differences between conditions (*F(2, 86) = 0.88, p > 0.05*). This finding was interesting since substituting or deleting PINTs material created minor artefacts within the audio. The original, unmanipulated version was found to be the most difficult to follow, possibly because this speaker uses a high frequency of PINTs (∼40% of his total speaking time). When comparing means of ease by L1, we found that the monolingual English group (M = 3.05 , SD = 1.29) and the L1 German group (M = 2.60, SD = 1.16) were not significantly different (*t(87) = 1.71, p > 0.05, d = 0.36*). These results indicate that the NNSs found the English-language lecturer as easy to understand as the monolingual NSs of English.

Linear regression models were tested with L1, condition, and the different questionnaire variables. The model with the best fit (lowest AIC) predicted total score with ease as the only fixed effect. Table 5.8 shows that with an ease value of 1, participants' total score was 4.84 (out of a total of 8) and that the higher the ease value, the higher the total score. Significant effects for all levels of ease were found, except for a value of 3. Importantly, an ease value of 5 improved participants' total score more than an ease value of 4 which improved more than an ease value of 2. The ease value of 3 did not follow this trend.

Correlation

Table 5.9 contains the Pearson correlations between total score and the questionnaire variables. No correlation was found between age or interest and total score. However, we found a weak correlation between ease and total score (*t(87) = 2.98, p < 0.01, r = 0.30*), and between preparation and total score (*t(87 = 2.77, p < 0.01, r = 0.28*).

5.2.5 Discussion

This experiment evaluated whether PINTs improved the recall of English-language lecture material for native and non-native listeners. We found that PINTs immediately preceding key information negatively impacted score. While we found that L1 did not influence the ease rating of the speaker, non-native students may encounter significant problems when listening to lectures, such as word recognition or with creating meaning (Kilbon, 2022). These issues are related to linguistic aptitude and awareness of the lecture material. This study found that monolingual English listeners tended to scored better in the "no PINTs" situation, while L1 German listeners tended to scored better with the original audio. This may be due to monolingual English listeners being above the threshold of needing the time-buying aspect of PINTs. Conversely, the L1 German listeners still benefited from the time-buying aspect of PINTs. Blau (1990) found that a base skill level might be required before the benefits of additional processing time from pauses can be seen, and importantly, that beyond a certain skill level pauses may no longer aid and, instead, be an irritant to the listener. Jacobs et al. (1988) found that pauses may increase comprehension, but only for advanced students. Jacobs et al. (1988) also found that increasing the duration of pauses has a ceiling and once above that ceiling, comprehension will decrease. Therefore, it is important to consider the impact of PINTs in environments where the recall of key information is crucial, such as educational settings, and for both native and non-native listeners.

This study investigated the influence of PINTs on recollection in an ecologically valid scenario with longer material lengths. We found PINTs to be detrimental to the recollection of upcoming content in a lecture scenario, which is contradictory to what previous studies have found. These results might differ from other studies due to the longer material lengths, or because there are many variables that are difficult to control in a real-world scenario. Additionally, this experiment treated all PINTs equally. In real-world scenarios, speakers have distinct PINTs profiles and often many PINTs will co-occur making it difficult, if not impossible, to evaluate individual PINTs separately. Furthermore, effects of both PINTs placement and frequency should be further explored. While non-native participants were asked for their test score, we were not able to get an accurate picture of the influence of PINTs for different skill levels. In this study most of the L1 German listeners were advanced in English. Future work should continue to evaluate longer material lengths, with a variety of language backgrounds and skill levels.

5.3 The Impact of Pause-Internal Phonetic Particles on Recall in Synthesized Lectures

5.3.1 Abstract

We studied the effect of pause-internal phonetic particles (PINTs) on recall for native and non-native listeners of English in a listening experiment with synthesized material that simulated a university lecture. Using a neural speech synthesizer trained on recorded lectures with PINTs annotations, we generated three distinct conditions: a base version, a "silence" version where non-silence PINTs were replaced with silence, and a "nopints" version where all PINTs, including silences, were removed. Half of the participants were informed they were listening to computer-generated audio, while the other half were told the audio was recorded with a poor-quality microphone. We found that neither the condition nor the participants' native language significantly affected their overall score, and the presence of PINTs before critical information had a negative effect on recall. This study highlights the importance of considering PINTs for educational purposes in speech synthesis systems.

5.3.2 Introduction

Research in speech synthesis for education is an important area of study that can yield benefits for both native speakers (NSs) and non-native speakers (NNSs). Kang et al. (2008) found that NNSs comprehended synthetic sentences more easily than synthetic words, and that ratings were dependent on the listener's comprehension level. Additionally, Kang et al. (2008) highlighted the need for investigating longer material lengths for speech synthesis. For example, when evaluating recall for larger contexts with real-world data, Elmers (2023) found no recall effect for PINTs for both native and non-native listeners. Specifically, Elmers (2023) used segments from English-language universities and found that PINTs reduced the recall for upcoming information. Therefore, we were interested in investigating how PINTs might influence synthesized speech in the same lecture setting, and whether their impact on recall was similar for both NSs and NNSs. Our present study replicated (Elmers, 2023), using the same textual material but with a novel approach: the material was synthesized instead of being natural speech, and two instruction sets were created. One instruction set informed participants they were listening to synthesized speech, while the other instruction set told participants that the audio was recorded with a poor-quality microphone. Our investigation focused on the impact of PINTs on the recall of sections of synthesized lectures.

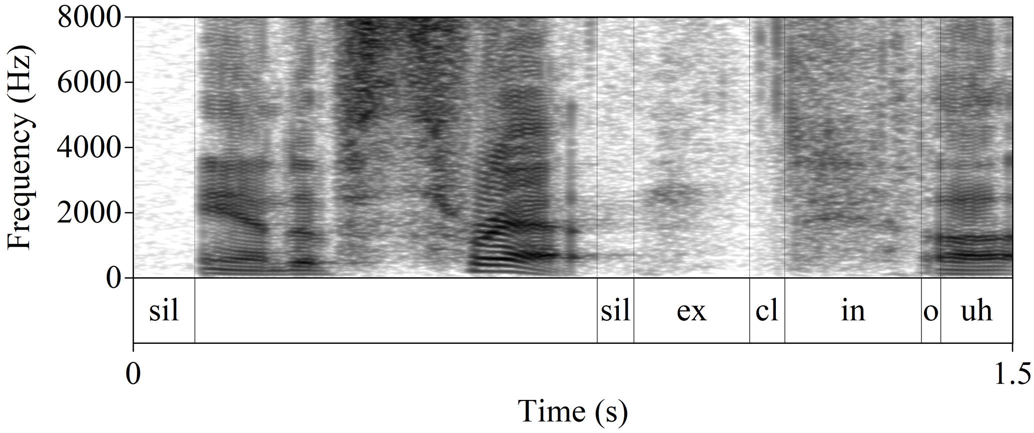


Figure 5.7: Spectrogram of original textual material. PINTs annotations: silence (sil), exhalation noise (ex), inhalation noise (in), filler particle (uh), tongue click (cl), and other (o).

5.3.3 Method

We sourced lecture material from Open Yale Courses (2007b), a collection of free and open access courses provided by Yale University. The speaker used for this study had a significant occurrence of PINTs during their lectures, accounting for 40% of their total lecture time. We annotated a subset of these lectures to train the speech synthesis model. Fig. 5.7 and Fig. 5.8 compare the same segment of speech for both the original and synthesized versions.

Stimuli

To generate the stimuli for the experiment, a neural text-to-speech (TTS) voice was created, using a method similar to the one described in Elmers et al. (2023). This TTS system was trained using a PyTorch implementation[6] of the sequence-to-sequence neural TTS engine Tacotron 2 (Shen et al., 2018). We used phoneme-level input for training and synthesis, which was obtained from the transcripts using the `g2p_en` package (Park & Kim, 2019). The training corpus was divided into segments, delineated by an inhalation breath on each end, where multiple breath groups were joined into utterances of at most 11 seconds. For the phonetic input, specific tokens for the different PINTs were added to allow exact prompted reproduction. The voice was trained for 70k iterations on top of the published read speech model, using 3 GPUs and a batch size of 28. The speech signal was decoded from the model output, using the neural vocoder HiFi-GAN (Kong et al., 2020). The published model was finetuned for 1.33M iterations on the corpus.

[6] https://github.com/NVIDIA/tacotron2

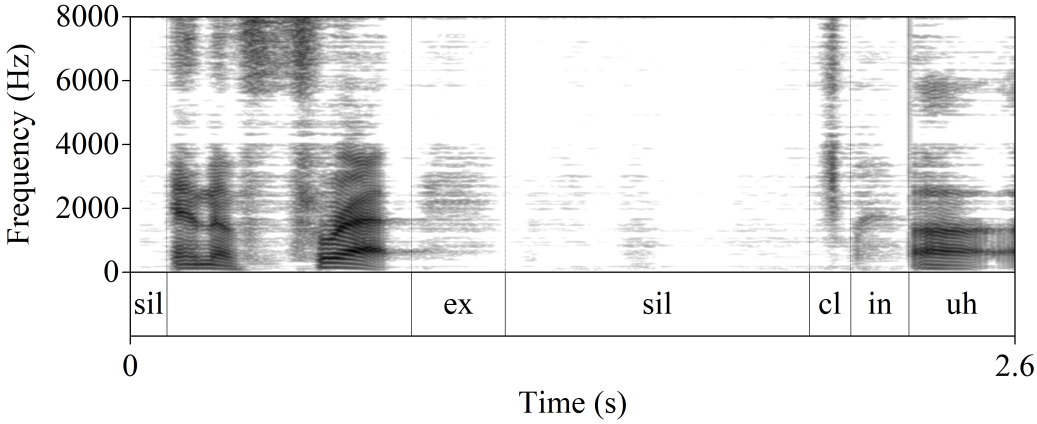


Figure 5.8: Spectrogram of synthetic speech textual material. PINTs annotations: silence (sil), exhalation noise (ex), inhalation noise (in), filler particles (uh), tongue click (cl).

The perceptual experiment material was generated one breath group at a time following Székely et al. (2019b, 2020), with each segment starting and ending with an inhalation. Since the synthesizer is non-deterministic, we synthesized multiple versions and chose the version that included fewest distortions or errors. Specifically, we avoided versions that included background noises, which occasionally occurred due to the varying audio quality of the speech corpus. After synthesizing all breath groups for each lecture segment, we concatenated the segments together using Praat (Boersma & Weenink, 2022).

The study we replicated included three different conditions for the audio stimuli. We created the same conditions using a Praat (Boersma & Weenink, 2022) script, which either removed or replaced the PINTs material. For the silence condition, the non-silence PINTs were replaced with silence taken from the audio that was adjusted to match the duration of the removed material. As a result, the silence condition maintains the same duration as the base condition. The "nopints" condition removed all acoustic pauses. Both the base and silence conditions provide the same processing time, whereas the "nopints" condition allows for less processing time (see Fig. 5.9).

Participants

This study used a web-based experiment, designed with the Labvanced[7] (Finger et al., 2017) platform, to present the audio stimuli and collect responses and questionnaire data from participants. Recruitment was carried out with the crowd-sourcing platform Prolific[8] (2014). A total of 180 participants were recruited, including 90 monolingual English participants (mean age 40 years, age range 20–75 years) and 90 L1

[7] Accessed via `https://www.labvanced.com/` on Apr. 12, 2023.

[8] Accessed via `https://www.prolific.co/` on Apr. 12, 2023.

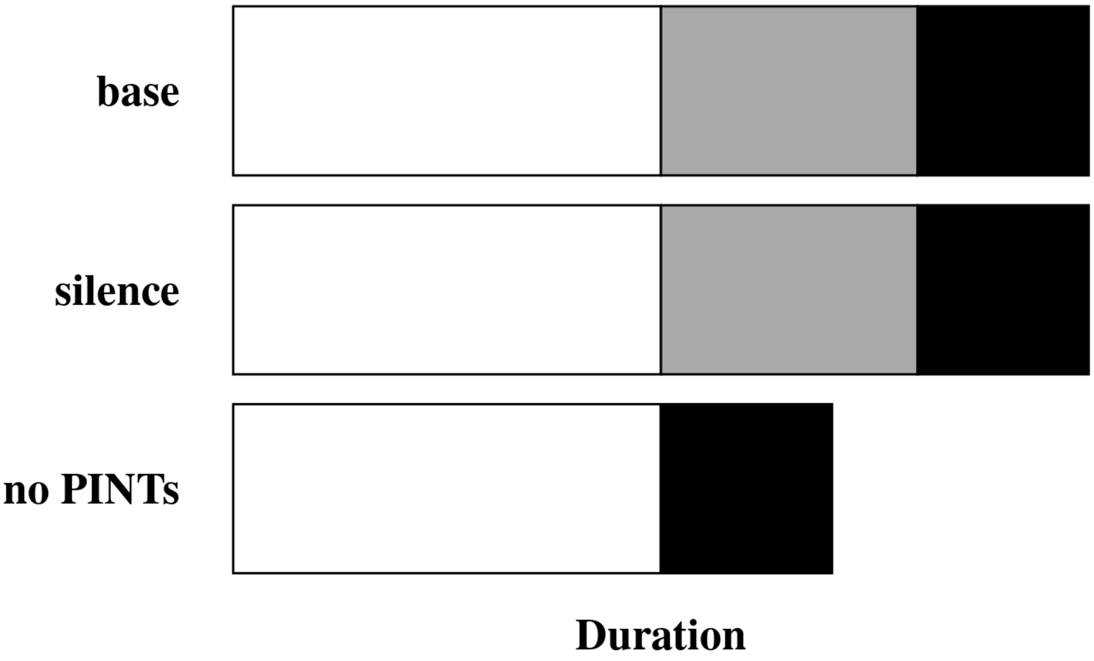


Figure 5.9: Duration schematic for the three conditions, with speech material in white, PINTs in grey, and speech material that includes the answer depicted in black.

German participants (mean age 33 years, age range 18–70 years), who received compensation for their participation. Three monolingual English participants and two L1 German participants self-reported hearing impairment and were excluded from the results.

Procedure

Participants listened to four sections of synthesized material, each approximately three minutes in length, and then answered two multiple-choice questions based on the content. The experimental material used the same textual content from Elmers (2023), which extracted sections from full-length lectures. The stimuli were carefully chosen to ensure that all information needed to answer the question was present in the audio, i.e., no prior knowledge was required to answer the questions. For each audio segment, one question was preceded by PINTs material, while the other was not. In the "nopints" condition, neither question contained PINTs material. An example question was, "According to Paul Fussell, what is the essential trope or rhetorical figure of World War One poetry?" Possible answers included: a) hyperbole, b) metaphor, c) oxymoron, d) irony. Participants only needed to understand the contents, as presented by the speaker, and did not require any encyclopedic or background knowledge to answer the questions. The study was balanced to ensure that the material preceding the first and second question was equally distributed, with or without PINTs.

Before beginning the experiment, participants received a set of instructions. They were informed that the study involved listening to four audio clips, each lasting approximately three minutes, and answering content-based questions immediately following each clip. The participants were instructed to use headphones and test their audio volume prior to commencing the experiment. There were told to not take any notes while listening to the audio clips. Half of the participants were informed that the audio was computer-generated audio, while the other half were informed that the recordings were made using a sub-optimal microphone, and contained some background noise. The latter group of participants were not informed that the audio was computer-generated. Participants were told that they would receive a score at the end of the experiment to incentivize them to perform well.

Following the listening task, participants were required to complete a questionnaire. The questionnaire included various demographic and language-related questions such as age, hearing impairment, L1, highest level of completed education, interest in the audio contents (1: very uninterested to 5: very interested), ease of following/understanding the speaker (1: very difficult to 5: very easy), and perception of the speaker's preparedness (1: very unprepared to 5: very prepared). The L1 German listeners were asked to provided their self-reported English proficiency and, if available, a test score. The total time for completing the listening task and questionnaire averaged between 15 to 20 minutes.

Each participant listened to one of three conditions. Specifically, one-third of the participants listened to four unmodified (i.e., "base" condition) audio clips, another one-third listened to four audio clips from the "silence" condition, and the remaining one-third listened to four clips from the "nopints" condition. The textual material was the same across all conditions, but the order of the four audio clips was randomized to prevent potential order effects. Participants saw "Listen closely!" displayed on their screen while the audio played, and each audio clip was played only once.

5.3.4 Results

R (R Core Team, 2021) (Version 4.1.2) was used to perform both descriptive and inferential statistics on the data[9]. Prior to analysis, the data was pre-processed using the dplyr (Wickham et al., 2023a) (Version 1.1.1), stringr (Wickham, 2022) (Version 1.5.0), and tidyr (Wickham et al., 2023b) (Version 1.3.0) packages. The homogeneity of variance assumption was assessed using Levene's test from the car (Fox & Weisberg, 2019) (Version 3.1.2) package. Cohen's d was calculated using the effsize (Torchiano, 2020) (Version 0.8.1) package. Post-hoc comparisons after conducting the Kruskal-Wallis rank sum test were performed using Dunn's Test from the dunn.test (Dinno, 2017) (Version 1.3.5) package. The statistical models were analyzed using either

[9] This project's data and scripts can be found at https://github.com/MikeyElmers/paper_ssw23.

Table 5.10: Duration information for the original and synthesized material. Values include all four passages. Total PINTs duration (PINTs dur) and total audio duration (total dur) are reported in seconds. Proportion (prop) is expressed as a percentage value out of 100%.

condition	PINTs dur	total dur	prop
original	314.45	761.48	41.29
synthesized	273.68	705.23	38.81

Table 5.11: Count information for the original and synthesized material. The values include all four passages. The following labels were investigated: silence (sil), inhalation (in), exhalation (ex), filler particles (uh) and (um), tongue click (cl), other (o).

condition	sil	in	ex	uh	um	cl	o
original	656	234	65	210	30	43	82
synthesized	288	232	9	212	30	45	2

linear regression or with binomial generalized linear mixed models (binomial GLMMs) implemented through the lme4 (Bates et al., 2015) (Version 1.1.32) and lmerTest (Kuznetsova et al., 2017) (Version 3.1.3) packages. The Akaike Information Criterion (AIC) (Akaike, 1973) was used to compare models and choose the best fit model with the lowest AIC. Visualizations were generated using ggplot2 (Wickham, 2016) (Version 3.4.2).

Duration and Count

We compared the count and duration information of the four synthesized passages to the original versions. The originals had a longer PINTs duration, a longer total duration, and a higher PINTs proportion (see Table 5.10). However, while the synthesized passages were shorter overall, the PINTs proportions were comparable. Both versions were similar to the speaker's overall PINTs profile proportion of 40%. Table 5.11 contains count information for the individual PINTs grouped by condition. The original material had more silences, exhalations, and "other" particles. Both versions had a similar number of inhalations, filler particles "uh" and "um", and tongue clicks.

L1 Comparison

Participant performance was determined by the number of correctly answered questions, with a potential maximum score of 8 (1 point for each correct answer). We compared total score means grouped by L1. The normality assumption for the dependent variable was violated, as determined by the Shapiro-Wilk Normality test (p

Table 5.12: Descriptive statistics for total score (possible maximum score of 8) for the different conditions. Participants were told that the audio was computer-generated in the conditions with the subscript “cg”.

condition	mean	median	sd	n
$silence_{cg}$	5.77	5	1.41	31
$nopints$	5.71	6	1.61	28
$base_{cg}$	5.61	6	1.64	28
$base$	5.45	5	1.78	29
$nopints_{cg}$	5.34	5	1.72	29
$silence$	4.77	5	1.68	30

$<$ 0.05). The assumption of equal variances was met, as verified by Levene’s test for homogeneity ($p > 0.05$). Since the sample sizes exceeded 30, a parametric test was used following the central limit theorem, despite the violation of normality. An independent samples t-test compared the mean scores of the monolingual English group (M = 5.31, SD = 1.67) and the L1 German group (M = 5.57, SD = 1.64). The findings revealed *no* significant difference between the two groups (t(173) = -1.03, p $>$ 0.05, d = -0.16), indicating that the L1 German group and monolingual English group performed similarly.

Condition and Group Comparison

Given that L1 was not a significant factor, we proceeded to investigate the effect of condition. Total scores grouped by condition are summarized in Table 5.12. Participants were informed that they were either listening to computer-generated audio (in half of the conditions) or that the speaker was using a poor-quality microphone (in the remaining half). Notably, the highest mean score was obtained in the silence condition where participants were told that the audio was computer-generated. The lowest mean score was observed in the silence condition where participants were told that the speaker used a poor-quality microphone. This pattern was not universal, but in general, when participants were told the audio was computer-generated, the total score was higher than when they were told the audio came from a poor-quality microphone, except for the “nopints” condition.

We conducted an analysis of variance (ANOVA) to investigate the mean differences in total score between conditions, with condition as a fixed factor. The normality assumption was violated in the following comparisons as indicated by the Shapiro-Wilk test, while Levene’s Test did not reveal any significant differences in variances across groups. Given that the sample sizes exceeded 30, satisfying the central limit theorem, we utilized parametric tests. The results did not indicate any significant effect of condition on total score (F(5, 169) = 1.50, $p > 0.05$). Post-hoc comparisons using pairwise t-tests with Bonferroni correction did not reveal any significant differences

Table 5.13: Descriptive statistics, including mean and standard deviation, based on whether the question material was immediately preceded by PINTs material.

precede	mean	sd
no	0.77	0.42
yes	0.58	0.49

between individual groups. However, the Bonferroni correction is conservative by nature. A two-sample t-test was conducted to compare the mean total score between the silence$_{cg}$ (M = 5.77, SD = 1.41) and the silence (M = 4.77, SD = 1.68) conditions. The results showed that the silence$_{cg}$ condition had a significantly higher total score (t(59) = 2.55, p <0.05, d = 0.65). We conducted additional analyses by grouping conditions based on audio type (i.e., base, silence, and nopints) and whether participants were informed that the audio was computer-generated (i.e., cg group and non-cg group). However, these analyses did not reveal any significant differences in mean total score when grouping by audio type (F(2, 172) = 0.44, $p > 0.05$) or by computer-generated instruction type (t(173) = 1.12, $p > 0.05$, d = 0.17).

Preceding Material

We explored the impact on recall based on whether the question material was preceded by PINTs. Table 5.13 demonstrates that answers immediately preceded by PINTs material had an overall lower score. The nopints and nopints$_{cg}$ conditions, which did not contain any PINTs information, were excluded from this analysis. Scores were out of 1, rather than 8, since the evaluation was done on a by-question basis rather than the subject's collective score. A score of 1 indicated a correct response, and 0 an incorrect response. Due to violations of both normality, as indicated by the Shapiro-Wilk test, and homogeneity of variances, as indicated by Levene's test, we used the non-parametric Wilcoxon rank sum test. Our analysis revealed a significant difference between the conditions where PINTs material immediately preceded the question, and those where it did not (W = 132396, $p < 0.001$). Specifically, the mean score for questions preceded by PINTs material was significantly lower than those without any preceding PINTs information. These results suggest that the presence of PINTs information immediately before key information, may have a detrimental effect on recall performance.

We utilized a binomial generalized linear mixed effects model to investigate the relationship between score (0 or 1) and preceding PINTs, L1, condition, and questionnaire variables. The model with the best fit, as determined by the lowest AIC, was $glmer(score \sim precede + interest + (1 \mid id), family = binomial)$. This model predicted score based on whether the answer was preceded by PINTs information and interest level as fixed effects, with the subject as a random effect. This model outperformed alternative models that incorporated L1 status, condition, or other questionnaire variables. Our findings indicated that both preceding PINTs and interest level significantly predicted score. The intercept was significant ($\beta = 0.84$, $p < 0.001$), and whether the answer was preceded by PINTs material had a significant negative effect ($\beta =$ -0.96, $p < 0.001$). Interest level had a significant effect on score, with the highest levels of interest being the strongest predictors ($\beta = 1.80$, $p < 0.001$ for interest level 5; $\beta = 0.98$, $p < 0.001$ for interest level 4; $\beta = 0.42$, $p > 0.05$ for interest level 3; $\beta = 0.38$, $p > 0.05$ for interest level 2). These results suggest that only the highest levels of interest (4 and 5) are associated with higher odds of a positive outcome on score. In summary, our findings indicate that when holding all other variables constant, the odds of a positive outcome on score decrease when PINTs precede, while the highest levels of interest are associated with higher odds of a positive outcome on score.

Interest

Participants rated their level of interest in the audio contents on a scale of 1 (very uninterested) to 5 (very interested), with a mean rating of 2.73. An independent samples t-test was conducted to compare the mean interest ratings between the monolingual English group (M = 2.70, SD = 1.22) and the L1 German group (M = 2.76, SD = 1.13). The results indicated no significant difference between the two groups (t(173) = -0.33, $p > 0.05$, d = -0.05).

Interest was further examined by condition (see Table 5.14). We used an ANOVA to investigate mean differences in interest between the different conditions. The results revealed a significant effect of condition on interest (F(5, 169) = 2.45, $p < 0.05$). Post-hoc pair-wise t-tests with Bonferroni correction showed a significant difference between the nopints condition and the base condition ($p < 0.05$). Both the L1 and condition comparisons violated normality, as indicated by the Shapiro-Wilk test. However, homogeneity of variances was maintained, as indicated by Levene's test. As a result, we opted for parametric tests since both comparisons satisfied the central limit theorem.

We used linear regression models to investigate the relationship between total score and L1, condition, and the questionnaire variables. The model with the lowest AIC included interest and instruction type (whether participant were told the audio was computer-generated) as fixed effects: $lm(total\ score \sim interest + cg)$. This model was statistically significant (F(5, 169) = 5.76, $p < 0.001$) and explained 14.56% of the variance in total score ($R^2 = 0.1456$). Our results indicated that interest level had a significant effect on total score, with the highest level of interest being the strongest predictor ($\beta = 2.42$, $p < 0.001$ for interest level 5; $\beta = 1.05$, $p < 0.01$ for interest level 4; $\beta = 0.42$, $p > 0.05$ for interest level 3; $\beta = 0.13$, $p > 0.05$ for interest level 2). However, instruction type did not have a significant effect on total score ($\beta = 0.38$, $p > 0.05$). The intercept was significant ($\beta = 4.72$, $p < 0.001$). The adjusted R-squared

Table 5.14: Mean interest values for the different conditions. Participants were told that the audio was computer-generated in the conditions with the subscript "cg".

condition	mean	sd
nopints	3.25	1.17
sil_{cg} is below; *sil*	2.90	1.03
sil_{cg}	2.77	1.09
$base_{cg}$	2.64	1.16
$nopints_{cg}$	2.59	1.18
base	2.24	1.27

Table 5.15: Pearson correlation coefficients for participants' total score and their questionnaire responses. Includes correlation information for all participants (Both), L1 German (DE), and monolingual English (EN).

participants	age	ease	interest	prep
Both	0.20	0.15	0.33	0.10
DE	0.15	0.07	0.25	−0.00
EN	0.30	0.24	0.39	0.19

of the model was 0.12, with a residual standard error of 1.55. These findings suggest that the level of interest in the audio content has a significant effect on total score, but instruction type does not have a significant effect on total score.

Correlation

Pearson correlations for participants' total score and their questionnaire responses are presented in Table 5.15. The questionnaire assessed how easy it was to comprehend the speaker, level of interest in the lecture content, and evaluated the speaker's preparedness. Ratings were made on a 5-point Likert scale. Age, ease, and preparedness were weakly correlated with total score, while interest was moderately correlated (t(173) = 4.56, $p < 0.001$, r = .33). The correlation between total score and interest was stronger for monolingual English participants than for L1 German participants. Overall, the higher the participant's interest level, the better their total score.

5.3.5 Discussion

Compared to the original version, the synthesized version struggled to generate short silences, often found adjacent to non-silence PINTs (i.e., edge silences), resulting in fewer silences. Similarly, the counts for exhalations and the "other" category also decreased in the synthesized version, possibly due to the scarcity of exhalations in the data. The "other" category was not included during the synthesizer's training due to the diversity of phenomena within the category. Nevertheless, the synthesizer generated some "other" labels without any explicit inclusion. Despite these differences, the synthesizer maintained similar counts for inhalations, filler particles "uh" and "um", and tongue clicks. However, the synthesizer occasionally produced multiple PINTs from a single label, leading to a higher count of "uh" and tongue clicks than was present in the original material. Overall, it was unexpected to find that the synthesized version closely modeled some of the counts, given that the output is not deterministic.

Our findings indicated that the participants' total score was not affected by their language background, whether they were monolingual English or L1 German, when listening to English-language content. This is a favorable outcome that suggests synthetic speech could be an effective equalizer for educational purposes, as both NSs and NNSs performed similarly. Additionally, we also observed significant differences in total score between the condition where participants were told the audio was synthesized (sil_{cg}), and the condition where participants were told the audio came from a poor-quality microphone (sil). This discrepancy may be due to participants being more lenient when they knew that the material was synthesized, as opposed to those who might be more critical assuming it was from a human speaker.

Our study found similar results to the study we replicated Elmers (2023), that PINTs preceding key information lowered recall. This highlights the need for future research to determine when PINTs can be beneficial for recall. However, neither study was able to replicate the benefits of PINTs observed in single sentence laboratory settings. It is possible that the PINTs profile of the speaker we used to train the TTS is an outlier. Lecture recordings from a speaker who uses PINTs to a lesser extent may reveal recall benefits. One limitation of using a single speaker is the difficulty in comparing how listeners perceive different realizations of the same PINTs. Moreover, it is challenging to isolate individual PINTs for analysis in spontaneous speech recordings, where many PINTs co-occur. In this experiment, we treated all PINTs equally, despite each PINT having different realizations that may impact recall differently.

Participants who rated their interest in the audio content as high (4 or 5) had a significantly higher total score, and this was reflected in a moderate correlation between interest and total score. Contrary to expectations, the instruction type did not have a significant effect on the total score, suggesting that whether the participants knew the material was synthesized or not, did not impact their performance. When comparing mean interest scores by condition, it was unexpected to find the "nopints" condition had a higher interest score than the "base" condition, despite the audio artefacts resulting from the removal of the pause material. One possible explanation is that, again, the speaker used for training the TTS model used too many PINTs during their lectures, which might have resulted in a less engaging experience.

For most levels of proficiency, Blau (1990) found that pauses helped comprehension more than speaking at a normal rate or artificially slowing the speaking rate. However, there is a threshold, beyond which pauses have a negative effect on comprehension (Jacobs et al., 1988; Blau, 1990). It is possible that the PINTs used in this study exceeded the threshold and became a detriment to recall. These results indicate that the impact of PINTs in synthesized speech should be carefully considered when recall of information is important, such as in education, for both native and non-native listeners.

Chapter 6

General Discussion

6.1 Recall

The first section of this dissertation focused on the influence of PINTs on recall in synthesized single-sentence laboratory contexts. Both silences (for digit recollection) and inhalations (for sentence recollection) showed a recall effect. In both experiments only the longer duration yielded the recall benefit. For the digit recollection experiment, we didn't evaluate different prosodic grouping structures often found when speaking a series of numbers. For example, a common grouping structure for 7-digit phone numbers is 3-2-2 (Baumann & Trouvain, 2001). Our results are unable to clarify possible interaction effects between prosodic groupings and PINTs insertion in synthesized speech. These results indicate the importance of evaluating multiple durations in a variety of contexts. For the sentence recollection experiment, our breath noises were chosen to be shorter than the versions used in Whalen et al. (1995). Our long breath noise duration was approximately the same duration as the short breath noise condition used by Whalen. Our results indicated that for synthetic speech, longer PINTs durations might be ideal for the purpose of recollection. Fraundorf & Watson (2011) posited three possible hypotheses for the durational benefit in recall: 1) a processing-time hypothesis where the longer durations give additional time to the listener, 2) an attention orienting hypothesis where the longer duration PINTs are better able to attract the attention of listeners, and 3) a predictive processing hypothesis where participants use the PINTs duration to make assumptions about the length of the upcoming material. Future work should continue to tease apart the relationship between duration and PINTs recollection, with a focus on the mechanism for the durational benefit.

Another area for future research is the experimental measurement. While both experiments used recall as their experimental measurement, digit recollection and sentence recollection are very different tasks. In the case of sentence recollection, participants were tasked with recalling the sentence verbatim. This isn't a trivial task for participants and, in addition to recall, required a focus on spelling and typing/writing their response. These additional requirements might have lowered the amount of focus available for recollection. This measurement was chosen because participants were recalling sentences. However, in the case of longer material lengths, this form of recollection isn't plausible. Longer material lengths may even help participants by increasing the naturalness of the situation and prevent boredom (Braunschweiler & Chen, 2013). Luckily, there are a number of interesting potential materials, such as audiobooks, or by incorporating dialogic elements between humans and conversational agents. This was a motivation for using a more realistic approach for measuring the effects of recall in lectures by evaluating the recall of key information. Another possible experimental measurement often used in perceptual studies is reaction time. While reaction time wasn't ideal for our purposes, it is an important measurement that can be used to evaluate different perceptual phenomena. Future work should also investigate the distinction between understanding and memorization. The sentence recollection task involved memorization, while the lecture recall experiments involved understanding the content. Our lecture recall experiments involved multiple-choice questions but there are other options, such as fill-in-the-blank or short-answer responses. Another way to measure understanding would be to have the participants summarize material using their own words, similar to Fraundorf & Watson (2011), in order to gauge their degree of understanding. Experimental measurements are important to establish during experimental design, and each brings different pros and cons. Future work should continue to evaluate the influence of PINTs using a variety of different experimental measurements, allowing researchers to investigate with more specificity.

For the sentence recollection experiment, we found that shorter sentences were recalled better than the longer sentences. However, we were unable to determine the exact length where sentence length transitions from being an easy recall to a difficult recall. This is compounded since recall is influenced by individual features. Furthermore, in this experiment we measured sentence length via number of words, since this experiment was a partial replication study of Whalen. However, this metric isn't very stable since sentences with a few words can still be quite lengthy when containing longer words. Future work should evaluate more stable metrics such as a speech timing unit (e.g., number of syllables).

Whalen found a learning effect where participants had increased recall in the second half of the stimuli. In our replication study we did not find any kind of learning effect. One possible explanation is the audio quality. Whalen used a formant synthesizer while our experiment generated stimuli with a concatenative synthesizer, possibly indicating that improvements to the audio quality for modern speech synthesis systems might be a cause for the lack of a learning effect. Another possibility is that listeners are more accustomed to hearing computer-generated audio, compared to when Whalen's study took place in the mid-1990s. We also incorporated a questionnaire where participants were asked about how often they hear computer-generated audio, such as conversational agents or in-car navigation systems. Most participants reported familiarity with computer-generated audio. Overall, future research should continue to investigate the relationship between PINTs and recall via duration, recollection measurements, length measurements, and learning effects.

6.2 Synthesis

The second section of this dissertation focused on the detection and synthesis of PINTs. During the detection experiment, the neural network (NN), convolutional neural network (CNN), and recurrent neural network (RNN) performed similarly. This contrasted with our hypothesis that the RNN would outperform the other models, since RNNs are able to account for temporal information. All models were able to easily classify the silence segments and inhalations, which were the most common annotations in the data. The models found moderate success with classifying exhalations, which was the next most common annotation type. Finally, all three models failed to classify FPs and clicks, which were the least frequent annotations within the data. Overall, these results suggest that the accurate classification of PINTs is more dependent on annotation quantity and quality than the model architecture. Moreover, all models began with a relatively high accuracy during training and improvement was minimal during the remaining epochs. Regarding misclassified segments, both inhalations and exhalations were sometimes wrongly classified as silence segments, which might be caused by the common adjacency of these PINTs. The models were unable to classify FPs, possibly because the models found them too similar to the speech category. Clicks were also unable to be classified and were generally misclassified as silence segments, possibly due to their shorter durations. The models were trained using mel-frequency cepstral coefficients (MFCCs) as input. Incorporating acoustic features or training on spectrogram images might improve the classification of PINTs and should be evaluated in future work.

Since we were unable to develop an automatic method for classifying multiple PINTs simultaneously, we proceeded with manual annotations for developing our speech synthesis systems. Manual annotations are a time-intensive process and limited our ability to model more than a single speaker. Therefore, we are unable to determine whether our results would transfer to other speakers who use different PINTs patterns. Future work should incorporate the PINTs patterns of multiple speakers to determine if certain patterns are more beneficial for certain tasks, like recall, than others.

We generated two TTS systems using the annotations from our single speaker: ControlledPINT and AutoPINT. The ControlledPINT system used our labeled annotations in the training data, while the AutoPINT model did not. The ControlledPINT system allowed the researcher to select a specific PINT and location. Conversely, the AutoPINT model inserts PINTs probabilistically. The ControlledPINT model produced more PINTs, both in terms of count and duration than the AutoPINT model. Importantly, the AutoPINT model showcased that even without labelled annotations, there was enough salient information to capture and model PINTs.

Both systems are, to the best of our knowledge, the first synthetic systems to produce tongue clicks, providing a technological and scientific contribution. Tongue clicks can display a number of different characteristics such as: indicating a new topic, word search, turn maintenance, backchanneling, stance marking, and repair (Ogden, 2013; Zellers, 2022). Additionally, the acoustic realizations for tongue clicks are highly variable (Trouvain & Malisz, 2016), indicating that the tongue clicks generated by our speech synthesis systems might not display our intended function of signaling a successful word search. Future work can further evaluate whether tongue clicks should be inserted or removed in order to create the desired meaning.

We compared the output of the ControlledPINT model to material from the same speaker that was not used during training. Our results showed that the ControlledPINT model produced counts similar to the original speech for inhalations, the filler particles "uh" and "um", and tongue clicks. Additionally, the ControlledPINT version sometimes produced a string of PINTs from a single label, resulting in a higher count for "uh" and tongue clicks than was used by the original speaker. The ControlledPINT model produced fewer silences overall compared the the original material. This was largely caused by the model struggling to generate edge silences, which are very short but also very frequent alongside other PINTs material. The ControlledPINT model also produced fewer exhalations, most likely due to lower frequency of annotations within the data.

Using the ControlledPINT model, we conducted a perceptual listening experiment to determine how inserted PINTs influenced certainty ratings for the audio. For the PINTs conditions, we expected the combinatory condition to elicit higher certainty scores since the inserted tongue click could indicate that the speaker had successfully accomplished their word search. However, this hypothesis was not verified in our experiment. The PINTs condition with the highest certainty scores was the long silence condition. A possible explanation is that the long silence condition was found less distracting or obtrusive than the FP or combinatory PINTs conditions. In this experiment all PINTs were evaluated equally. However, each PINT has different realizations that may affect perceived speaker certainty. This is a complex problem, especially since many PINTs co-occur, creating difficulties when trying to evaluate the effects of individual PINTs. Going forward it will be important to continue to evaluate a variety of PINTs, in a multitude of contexts, to further elaborate on the effects of different realizations. This experiment was also unable to verify possible influences of dialect, age, and gender. Future work should continue to evaluate how these factors interact with PINTs material. The experiment from our TTS system showcased that stimuli generated from speech synthesis systems has potential for providing greater experimental control. Moreover, as King (2015) notes, experimental design in the speech and hearing sciences has struggled to incorporate modern technologies, resulting in researchers compromising their materials. While generative modeling won't replace other research methods, it is an important and useful tool.

6.3 Lectures

The final section consisted of three experiments. In the first experiment we compared the PINTs usage of Yale University lectures to the TOEFL iBT, a popular English-language proficiency exam with a lecture listening component. Our initial hypothesis was that since lectures are a high proficiency speech style, that we would find minimal PINTs both for the Yale lectures and TOEFL lectures. Instead, we found approximately 1/3 of the total time for Yale lectures included PINTs material, and approximately 1/5 of the total time of the TOEFL lecture listening section included PINTs. These findings stress the importance of evaluating the influence of PINTs, especially for the recall of key information. Moreover, these findings illustrate that by not including PINTs in synthesized lectures, approximately 1/3 of total material is being lost and ignored. Often it is assumed that silences contribute to a majority of pause durations. However, we found that approximately 2/3 of the pause duration were PINTs other than silence. This experiment provided a baseline for the PINTs usage in a lecture style, which was important to developing a model that synthesizes PINTs with a lecture style. Future work should continue to evaluate the PINTs distribution of additional speakers in the lecture style, and in styles other than lectures. Non-native speakers are a major demographic for both the TOEFL iBT and university lecture courses. Therefore, it is important to extend the focus on the perceptual effect of PINTs beyond native speakers.

The second and third experiments investigated the influence of PINTs on recall for English-language lecture material. While experiment two used natural speech, experiment three used synthesized speech. Overall, our results indicated that recall was negatively impacted if the key information was immediately preceded by PINTs material. An important aspect of experiment two and three was the focus on ecologically valid scenarios that utilized materials longer than a single-sentence, in this case lecture segments. The recall effect of PINTs has been found for both natural and synthesized speech for single-sentence lengths, however, we were unable to replicate these results with lecture segments both for natural and synthesized speech. The results for the lecture segments could differ, because of the longer material lengths, or possibly because there are a variety of variables that are difficult to control in real-world contexts. Future work should continue to investigate a variety of longer material lengths and languages beyond English.

In both experiments we also captured questionnaire information. The questionnaire for experiment three revealed some interesting effects for interest. Participants who rated their interest levels as high for the audio content had a significantly higher recall. When investigating the relationship between interest and the stimuli condition, certain trends emerged. The “no PINTs” condition had higher interest scores than the “base” condition, even though the audio incorporated minor artefacts that resulted from removing the PINTs material. This might be caused by the way some audio media is digested. For example, audio material can be listened to at 2x speed, which significantly reduces perceivable PINTs. Therefore, if it is becoming more common to listen to content-based audio at higher speeds, then the “no PINTs” condition might be more interesting due to it’s shortened length. Another avenue for future research is evaluating the perceptual effects of different PINTs profiles. In our experiment we used a single speaker, so it is possible our chosen speaker is an outlier in PINTs usage. Another possibility is that the speaker used too many PINTs found in the “base” condition, which resulted in a less interesting style for listeners. Therefore, it would be beneficial to evaluate, in both natural and synthetic speech, other PINTs profiles. Future work should continue to evaluate how PINTs, in a variety of settings and styles, influences the interest levels of listeners.

Experiment three incorporated an instruction type condition where half of the participants were told that they were listening to computer-generated audio, while the other half were told they were listening to audio produced using a poor-quality microphone. Overall, the instruction type did not have a significant effect on recall, indicating that participants did not adjust their expectations based on whether or not they knew the audio was synthesized. However, we did find significant differences between the interaction of the condition and instruction type for the silence condition. When participants heard the silence condition, and were told that the audio was computer-generated, their recall score was significantly higher than when they heard the silence condition and were told that the audio came from a poor-quality microphone. Participants might have been more tolerable with the longer silences when they thought it was produced by a computer than a human. Since interactions with conversational systems is becoming more frequent in everyday life, it is important to understand how interlocutors adjust their perceptions based on whether they think they are conversing with a live or synthetic agent.

Since many students at university are non-native speakers of English, we evaluated both native and non-native speakers in experiments two and three. We did not find a significant difference between the recall scores for native and non-native participants in either experiment. In both studies, we used L1 German participants who had an intermediate or advanced level of English. A limitation of our experiments was that we couldn’t evaluate how skill level and the recall effect of PINTs interact. Our results demonstrated that future research should continue to determine whether PINTs can aid in recall, and evaluate possible interaction effects for native or non-native speakers.

Non-native students can experience a variety of problems while listening to lectures, such as word recognition or meaning creation (Kilbon, 2022). Blau (1990) found that pauses aided in comprehension for most proficiency levels, more than using a normal speaking rate, or artificially slowing the speaking rate. However, the benefits have a limit, and beyond this threshold pauses were found to have a negative effect on comprehension (Jacobs et al., 1988; Blau, 1990). It is possible that the PINTs stimuli incorporated in these studies exceeded this threshold, and became a detriment to recall. For example, Betz et al. (2015) noted that we are still unsure about the maximum acceptable duration of silences in TTS systems before they become disruptive, and Rose (2013) indicated that silence rate and duration have implications in the evaluation of L2 proficiency.

It's important to consider the influence of technologies in educational settings. For example, natural language processing can be applied to evaluate students while learning a language (Meurers, 2020). Furthermore, Adell et al. (2012) argues that TTS systems should emulate spoken speech, rather than read speech. Ewer (1974) stresses the need to create realistic lectures by including PINTs, as well, as gestures, glances, and other hesitation phenomena. Similarly, Wang et al. (2021) concluded that it is important to combine the synthesis of speech and gesture. The relationship and coordination of PINTs and gesture is not evaluated in this work but is certainly an important area for future research. Mazzocconi et al. (2022) argues the inclusion of laughter is also crucial for affective aspects, natural language understanding, and pragmatic reasoning. Additionally, Betz et al. (2016) advocates for the inclusion of lenthenings as a strategy for providing additional time.

Synthetic speech could be a potential equalizer for educational purposes, evident from our finding that synthetic speech stimuli didn't create additional recall issues for non-native listeners. This is an important finding, considering the wealth of opportunities synthetic speech provides to educational applications. This work incorporated a variety of methods for evaluating synthetic speech such as: digit recall, sentence recall, lecture segment recall, and questionnaires. Many evaluators of TTS systems still focus on mean opinion scores, and while these are valuable, it is important to continue to evaluate a variety of objective and subjective metrics. Mendelson & Aylett (2017) stressed the importance of evaluating TTS with creative implementations that incorporate appropriate use cases. Moreover, Wagner & Betz (2017) concluded that TTS evaluations can be greatly improved by designing engaging situations that incorporate both subjective and objective measurements. The present work has shown one possible example of designing TTS for educational purposes that incorporates engaging impressionistic and behavioral metrics within the experimental design.

Chapter 7

Conclusion

The findings from this body of work contribute to a deeper understanding of the perceptual benefits of pause-internal phonetic particles (PINTs) in both natural and synthetic speech. While previous studies have generally only investigated PINTs in single-sentence laboratory experiments, this work has also investigated longer material lengths from semi-spontaneous sources.

PINTs can have significant effects in the processing of audio materials, especially in regards to recall. As speech synthesis systems strive to be more human-like they will continue to include PINTs. Going forward it will be important to understand and evaluate how these particles influence the perception of synthesized speech. This work initially investigated the perceptual effects when PINTs are inserted into single-sentence environments rendered by a text-to-speech (TTS) system. By analyzing the influences of PINTs in these shorter contexts, we found that PINTs improved recall. However, the TTS systems used for these experiments were limited in terms of PINTs control. Most TTS systems could not render any PINTs material. Systems that were able to render, could only implement silence or an inhalation, but not both. Moreover, these systems did not offer the kind of fine-tuned control that is ideal for experimental design. This led to the development of two custom TTS systems that could render PINTs: ControlledPINT and AutoPINT. The two systems differed only on their training data. The ControlledPINT system included lexical tokens for annotated PINTs material, allowing the researcher to specify the type and location of PINTs material. The AutoPINT system did not include lexical tokens for the annotated PINTs material, and rendered PINTs material entirely based on probabilistic modeling. To the best of our knowledge, these are the first models to produce tongue clicks. The ControlledPINT model was used to further evaluate the perceptual effects of PINTs on recall. However, instead of using single-sentence environments, we evaluated three-minute lecture segments that were extracted from full university lectures. Using both natural speech and the synthesized output from the ControlledPINT system, we evaluated the effect of PINTs on recall for native and non-native speakers of English. For both the natural and synthetic speech, PINTs negatively affected the recall of upcoming material. These findings were in contrast to the recall improvement found in single-sentence contexts. Importantly, the influence of PINTs, in both the natural and synthetic setting, did not impact non-native listeners more than native listeners. Through this research, we contributed to the broader scientific understanding of PINTs in synthetic speech, their influence on recall, and possible educational effects.

Chapter 3 consisted of two experiments investigating the effects of PINTs on recall in synthesized speech. Both experiments evaluated single sentences contexts. The first experiment investigated digit recollection. Participants heard a 7-digit number generated by a TTS system and were tasked with recalling a sequence of three adjacent digits. Some of the stimuli inserted a 200 ms or 500 ms silence before one of the digits. Results showed that recall accuracy improved after the inserted silent segment. We found a significant effect for the 500 ms silence condition, but not for the 200 ms condition. The second experiment investigated sentence recollection and was a partial replication of Whalen et al. (1995). Sentences were rendered using a concatenative synthesis system that included sentence initial inhalation breath noises. Three conditions of inhalation duration were evaluated: 0 ms (no inhalation), 300 ms (short duration inhalation), and 600 ms (long duration inhalation). Results indicated that sentences immediately preceded by the 600 ms inhalation were recalled better than sentences that were preceded by the 0 ms or 300 ms conditions. In summary, both experiments found recall benefits for PINTs in synthesized speech. The TTS systems used for both of these experiments did not provide much control in terms of PINTs. Therefore, the next step was to develop a custom TTS system that provided additional PINTs control for stimuli generation.

Chapter 4 consisted of two experiments with the goal of developing a bespoke speech synthesis system capable of generating PINTs. The first experiment sought to automate the process of labeling and annotating PINTs via evaluating different machine learning architectures. Using an annotated subset from a German spontaneous speech corpus, mel-frequency cepstral coefficients were extracted as inputs for modeling PINTs. Three different model architectures were compared: a general neural network (NN), a convolutional neural network (CNN), and a recurrent neural network (RNN). Models used the same hyperparameters, number of layers, and neurons per layer. Therefore, the model architecture was the main point of comparison. We hypothesized that the RNN would outperform the NN and CNN, since it is better able to handle temporal information. However, this was not the case, as all the models performed similarly. All models were able to detect silent segments and inhalation breath noises, the most common PINTs in the data. However, all models found moderate success when classifying exhalation breath noises, and failed to detect filler particles and tongue clicks. These results indicated that modeling multiple PINTs simultaneously doesn't always improve the classification accuracy of other surrounding PINTs. Moreover, accurate classification appeared to be more dependent on annotation quantity and quality than the model architecture. Since none of the classifiers were able to adequately detect all PINTs of interest, we proceeded to manually annotate a subset of university lectures. In the second experiment we developed two neural TTS systems: ControlledPINT and AutoPINT. The ControlledPINT system used labelled data for training, while the AutoPINT model rendered PINTs material probabilistically. When compared to the natural material, both systems produced less PINTs material for count and duration. However, the total PINTs proportion of the ControlledPINT system was much closer to the natural version than the AutoPINT version. Next, we used stimuli generated by the ControlledPINT system in a perceptual experiment to evaluate certainty scores. Since the ControlledPINT version allowed control over the type of PINT and location, we created four conditions: a PINTless (base) condition, a long silence condition, a filler particle condition, and a combinatory condition that included multiple PINTs. Our hypotheses was that the PINTless condition would be rated highest, and that the combinatory condition would be rated second highest. Results indicated that the PINTless condition was rated as the most certain, however, the combinatory condition was found to be the least certain.

Chapter 5 used the ControlledPINT model to, again, evaluate the influence of PINTs on recall in synthesized speech. This time the focus evolved past single-sentence contexts and onto three-minute segments excised from full length English-language university lectures. In the first experiment, we evaluated natural speech to develop a baseline. The goal was to incorporate a realistic educational setting where the usage of PINTs might have an important outcome, for example, quizzes in a university course. Additionally, we evaluated both native speakers of English and non-native speakers of English (L1 German), in order to determine if the effects of PINTs differ between the two groups. Three conditions were developed: an unmanipulated original version, a “silence” version where all non-silence PINTs were replaced with silence of the same duration, and a “no PINTs” version where PINTs were removed altogether. In both the original and “silence” conditions, half of the key information was preceded by PINTs material. The key information was crucial for the participant to correctly answer the question after listening to the audio contents. This paradigm allowed us to see if the the PINTs material immediately preceding the key information was assisting the listener in the recall of the aforementioned key information. The results indicated that key information immediately proceeded by PINTs material was less likely to be recalled. Additionally, both native and non-native speakers of English performed similarly. In experiment two, we used the same experimental paradigm as experiment one. However, experiment two used synthesized speech instead of the natural speech. The results were consistent with experiment one, in that participants’ recall score was not effect by their language background. Again, we found that the presence of PINTs information immediately preceding key information had a detrimental effect on recall. These findings differ from the recall benefits found in single-sentence environments.

Collectively, this work contributes to our understanding of the recall effect of PINTs, PINTs in speech synthesis systems, and PINTs in educational settings. This body of work has shown the influence that PINTs can have in a variety of settings. We've shown that PINTs improve recall in single-sentence contexts. Conversely, PINTs did not improve recall in longer segments of speech. These studies exemplify the large amount of diversity found within the application of PINTs material, and that nuance is critical to their evaluation. The conflicting findings between the single-sentence recall experiments and the longer context recall experiments indicate that there is still a large, and exciting, area to explore within this space.

List of Figures

List of Tables

Bibliography

Adell, J., Bonafonte, A., & Escudero, D. (2007). Filled pauses in speech synthesis: towards conversational speech. In *Text, Speech and Dialogue: 10th International Conference* (pp. 358–365). URL: https://doi.org/10.1007/978-3-540-74628-7_47.

Adell, J., Bonafonte, A., & Escudero, D. (2010). Synthesis of filled pauses based on a disfluent speech model. In *ICASSP 2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4810–4813). URL: https://doi.org/10.1109/ICASSP.2010.5495136.

Adell, J., Bonafonte, A., & Escudero-Mancebo, D. (2008). On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms. In *Proc. Interspeech 2008* (pp. 2278–2281). URL: https://doi.org/10.21437/Interspeech.2008-559.

Adell, J., Escudero, D., & Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, *54*, 459–476. URL: https://doi.org/10.1016/j.specom.2011.10.010.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory* (pp. 267–281). URL: https://doi.org/10.1007/978-1-4612-1694-0_15.

Amazon Web Services (2016). Amazon Polly. URL: https://aws.amazon.com/polly/ accessed: 10.01.2021.

Andersson, S., Georgila, K., Traum, D., Aylett, M., & Clark, R. A. J. (2010). Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Proc. Speech Prosody 2010*. URL: https://www.isca-speech.org/archive/speechprosody_2010/andersson10_speechprosody.html.

Audhkhasi, K., Kandhway, K., Deshmukh, O. D., & Verma, A. (2009). Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In *ICASSP 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4857–4860). URL: https://doi.org/10.1109/ICASSP.2009.4960719.

Aylett, M., Carmantini, A., & Braude, D. (2022). Why is my social robot so slow? How a conversational listener can revolutionize turn-taking. In *Proc. IROS 2022*.

Aylett, M. P., Vinciarelli, A., & Wester, M. (2020). Speech synthesis for the generation of artificial personality. *IEEE Transactions on Affective Computing*, *11*, 361–372. URL: https://doi.org/10.1109/TAFFC.2017.2763134.

Baggia, P., Bagshaw, P., Bodell, M., Huang, D. Z., Xiaoyan, L., McGlashan, S., Tao, J., Jun, Y., Fang, H., Kang, Y. et al. (2010). Speech synthesis markup language (ssml) version 1.1. *World Wide Web Consortium, Recommendation REC-speechsynthesis11-20100907*, . URL: https://www.w3.org/TR/speech-synthesis11.

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*, 441–455. URL: https://doi.org/10.1080/01690960903047122.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. URL: https://doi.org/10.18637/jss.v067.i01.

Baumann, S., & Trouvain, J. (2001). On the prosody of German telephone numbers. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (pp. 557–560). URL: https://doi.org/10.21437/Eurospeech.2001-149.

Belz, M., & Klapi, M. (2013). Pauses following fillers in l1 and l2 german map task dialogues. In *Proc. Disfluency in Spontaneous Speech (DiSS '13)* (pp. 9–12). URL: https://www.isca-speech.org/archive/diss_2013/belz13_diss.html.

Belz, M., & Trouvain, J. (2019). Are 'silent' pause always silent? In *Proc. 19th International Congress of Phonetic Sciences (ICPhS '19)* (pp. 2744–2748). Melbourne. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/.

Benus, S., Enos, F., Hirschberg, J., & Shriberg, E. (2006). Pauses in deceptive speech. In *Proc. Speech Prosody 2006* (p. 212). URL: https://www.isca-speech.org/archive/speechprosody_2006/benus06_speechprosody.html.

Betz, S., & Gambino, S. L. (2016). Are we all disfluent in our own special way and should dialogue systems also be? In *Proc. 27th Conference Elektronische Sprachsignalverarbeitung (ESSV '16)* (pp. 168–174). URL: https://www.essv.de/paper.php?id=337.

Betz, S., & Kosmala, L. (2019). Fill the silence! basics for modeling hesitation. In *Proc. Disfluency in Spontaneous Speech (DiSS '19)* (pp. 11–14). URL: https://doi.org/10.21862/diss-09-004-betz-kosm.

Betz, S., Wagner, P., & Schlangen, D. (2015). Micro-structure of disfluencies: basics for conversational speech synthesis. In *Proc. Interspeech 2015* (pp. 2222–2226). URL: https://doi.org/10.21437/Interspeech.2015-129.

Betz, S., Wagner, P., & Voße, J. (2016). Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. *Tagungsband Der 12. Tagung Phonetik Und Phonologie Im Deutschsprachigen Raum*, .

Blau, E. K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL quarterly*, *24*, 746–753. URL: https://doi.org/10.2307/3587129.

de Boer, M. M., Quené, H., & Heeren, W. F. L. (2022). Long-term within-speaker consistency of filled pauses in native and non-native speech. *JASA Express Letters*, *2*. URL: https://doi.org/10.1121/10.0009598.

Boersma, P., & Weenink, D. (2022). Praat: doing phonetics by computer. URL: http://www.praat.org.

Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning*, *64*, 579–614. URL: https://doi.org/10.1111/lang.12067.

Braun, A., & Rosin, A. (2015). On the speaker specificity of hesitation markers. In *Proc. 18th International Congress of Phonetic Sciences (ICPhS '15)*. Glasgow. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0731.pdf.

Braunschweiler, N., & Chen, L. (2013). Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. In *Proc. 8th ISCA Workshop on Speech Synthesis (SSW 8)* (pp. 1–6). Barcelona. URL: https://www.isca-speech.org/archive/ssw_2013/braunschweiler13_ssw.html.

Braunschweiler, N., & Maia, R. (2016). Pause prediction from text for speech synthesis with user-definable pause insertion likelihood threshold. In *Proc. Interspeech 2016* (pp. 3191–3195). URL: https://doi.org/10.21437/Interspeech.2016-752.

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Proc. Speech Prosody 2002* (pp. 199–202). URL: https://www.isca-speech.org/archive/speechprosody_2002/campione02_speechprosody.html.

Carlmeyer, B., Betz, S., Wagner, P., Wrede, B., & Schlangen, D. (2018). The hesitating robot - implementation and first impressions. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)* (pp. 77–78). URL: https://doi.org/10.1145/3173386.3176992.

Chollet, F. et al. (2015). Keras. URL: https://github.com/fchollet/keras.

Christensen, R. H. B. (2022). Ordinal—regression models for ordinal data. URL: https://CRAN.R-project.org/package=ordinal r package version 2022.11-16.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111. URL: https://doi.org/10.1016/S0010-0277(02)00017-3.

Condron, S., Clarke, G., Klementiev, A., Morse-Kopp, D., Parry, J., & Palaz, D. (2021). Non-verbal vocalisation and laughter detection using sequence-to-sequence models and multi-label training. In *Proc. Interspeech 2021* (pp. 2506–2510). URL: https://doi.org/10.21437/Interspeech.2021-1159.

Coppock, H., Gaskell, A., Tzirakis, P., Baird, A., Jones, L., & Schuller, B. (2021). End-to-end convolutional neural network enables covid-19 detection from breath and cough audio: a pilot study. *BMJ Innovations*, *7*, 356–362. URL: https://doi.org/10.1136/bmjinnov-2021-000668.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668. URL: https://doi.org/10.1016/j.cognition.2006.10.010.

Dall, R., Tomalin, M., & Wester, M. (2016). Synthesising filled pauses: Representation and datamixing. In *Proc. 9th ISCA Workshop on Speech Synthesis (SSW 9)* (pp. 7–13). URL: https://doi.org/10.21437/SSW.2016-2.

Dall, R., Tomalin, M., Wester, M., Byrne, W., & King, S. (2014a). Investigating automatic & human filled pause insertion for speech synthesis. In *Proc. Interspeech 2014* (pp. 51–55). URL: https://doi.org/10.21437/Interspeech.2014-11.

Dall, R., Wester, M., & Corley, M. (2014b). The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech. In *Proc. Interspeech 2014* (pp. 56–60). URL: https://doi.org/10.21437/Interspeech.2014-12.

Di Napoli, J. (2020). Filled pauses and prolongations in Roman Italian task-oriented dialogue. In *Proc. Laughter and Other Non-Verbal Vocalisations Workshop ('20)* (pp. 24–27). URL: https://doi.org/10.4119/lw2020-915.

Dinno, A. (2017). *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*. URL: https://CRAN.R-project.org/package=dunn.test r package version 1.3.5.

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech*, *25*, 11–28. URL: https://doi.org/10.1177/002383098202500102.

Echevarría, R. G. (2009). Cervantes' Don Quixote (Yale university: Open Yale courses). https://oyc.yale.edu/spanish-and-portuguese/span-300. License: Creative Commons BY-NC-SA.

Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. In *Proc. ITRW on Disfluency in Spontaneous Speech (DiSS '01)* (pp. 5–8). URL: https://www.isca-speech.org/archive/diss_2001/eklund01_diss.html.

Elmers, M. (2022). Comparing detection methods for pause-internal particles. In *Proc. 33rd Conference Elektronische Sprachsignalverarbeitung (ESSV '22)* (pp. 204–211). URL: https://www.essv.de/paper.php?id=1160.

Elmers, M. (2023). Pause particles influencing recollection in lectures. In *Proc. 20th International Congress of Phonetic Sciences (ICPhS '23)* (pp. 37–41). Prague. URL: https://guarant.cz/icphs2023/85.pdf.

Elmers, M., O'Mahony, J., & Székely, É. (2023). Synthesis after a couple PINTs: Investigating the role of pause-internal phonetic particles in speech synthesis and perception. In *Proc. Interspeech 2023* (pp. 4843–4847). URL: https://doi.org/10.21437/Interspeech.2023-2178.

Elmers, M., & Székely, É. (2023). The impact of pause-internal phonetic particles on recall in synthesized lectures. In *Proc. 12th ISCA Speech Synthesis Workshop (SSW 12)* (pp. 204–210). URL: https://doi.org/10.21437/SSW.2023-32.

Elmers, M., & Trouvain, J. (2022). Pause-internal particles in university lectures. Poster presentation at 18th Phonetik & Phonologie (P&P '22). URL: https://mikeyelmers.github.io/publications/elmers_pp22_poster.pdf.

Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. (2021a). Evaluating the effect of pauses on number recollection in synthesized speech. In *Proc. 32nd Conference Elektronische Sprachsignalverarbeitung (ESSV '21)* (pp. 289–295). URL: https://www.essv.de/paper.php?id=1131.

Elmers, M., Werner, R., Muhlack, B., Möbius, B., & Trouvain, J. (2021b). Take a breath: Respiratory sounds improve recollection in synthetic speech. In *Proc. Interspeech 2021* (pp. 3196–3200). URL: https://doi.org/10.21437/Interspeech.2021-1496.

Ewer, J. (1974). Note-taking training for non-English-speaking students of science & technology. *RELC Journal*, *5*, 41–49. URL: https://doi.org/10.1177/003368827400500104.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*, 313–326. URL: https://doi.org/10.1111/j.1467-1770.1987.tb00573.x.

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). Labvanced: a unified JavaScript framework for online studies. In *International Conference on Computational Social Science (Cologne)*.

Flowerdew, J., & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes*, *16*, 27–46. URL: https://doi.org/10.1016/S0889-4906(96)00030-0.

Flowerdew, J., & Tauroza, S. (1995). The effect of discourse markers on second language lecture comprehension. *Studies in second language acquisition*, *17*, 435–458. URL: https://doi.org/10.1017/S0272263100014406.

Fors, K. L. (2015). *Production and perception of pauses in speech*. Ph.D. thesis University of Gothenburg. URL: https://gupea.ub.gu.se/handle/2077/39346.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*. (3rd ed.). Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*, 161–175. URL: https://doi.org/10.1016/j.jml.2011.03.004.

Fruehwald, J. (2016). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, *22*, 41–49. URL: https://repository.upenn.edu/pwpl/vol22/iss2/6/.

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*, 29–47. URL: https://doi.org/10.1016/j.wocn.2012.08.007.

Fukuda, T., Ichikawa, O., & Nishimura, M. (2018). Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition. *Speech Communication*, *98*, 95–103. URL: https://doi.org/10.1016/j.specom.2018.01.008.

Garcia, A., Collery, M., Miloulis, V., & Malisz, Z. (2018). Classification and clustering of clicks, breathing and silences within speech pauses. In *Proc. 5th Laughter Workshop* (pp. 6–9).

Germesin, S., Becker, T., & Poller, P. (2008). Domain-specific classification methods for disfluency detection. In *Proc. Interspeech 2008* (pp. 2518–2521). URL: https://doi.org/10.21437/Interspeech.2008-624.

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*, 96–106. URL: https://doi.org/10.1080/17470215808416261.

Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Language and Speech*, *4*, 232–237. URL: https://doi.org/10.1177/002383096100400405.

Goto, M., Itou, K., & Hayamizu, S. (1999). A real-time filled pause detection system for spontaneous speech recognition. In *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* (pp. 227–230). URL: https://doi.org/10.21437/Eurospeech.1999-60.

Gustafson, J., Beskow, J., & Székely, É. (2021). Personality in the mix - investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)* (pp. 48–53). URL: https://doi.org/10.21437/SSW.2021-9.

Hammer, L. (2007a). Modern poetry (Yale university: Open Yale courses). https://oyc.yale.edu/english/engl-310. License: Creative Commons BY-NC-SA.

Hammer, L. (2007b). Open Yale courses. https://oyc.yale.edu/. License: Creative Commons BY-NC-SA.

Henter, G. E., Ronanki, S., Watts, O., Wester, M., Wu, Z., & King, S. (2016). Robust tts duration modelling using dnns. In *ICASSP 2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5130–5134). URL: https://doi.org/10.1109/ICASSP.2016.7472655.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, *23*, 99–132. URL: https://doi.org/10.1558/ijsll.v23i1.29874.

Ito, K., & Johnson, L. (2017). The LJ speech dataset. URL: https://keithito.com/LJ-Speech-Dataset/.

Jacobs, G., Chuawanlee, W., Itoga, B. K., Sakumoto, D., Saka, S., & Meehan, K. A. (1988). The effect of pausing on listening comprehension. In *The Eighth Second Language Research Forum* (pp. 1–19). Honolulu, HI: ERIC. URL: https://eric.ed.gov/?id=ED304018.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech Language and the Law*, *12*, 174–213. URL: https://doi.org/10.1558/sll.2005.12.2.174.

Kang, M., Kashiwagi, H., Treviranus, J., & Kaburagi, M. (2008). Synthetic speech in foreign language learning: an evaluation by learners. *International Journal of Speech Technology*, *11*, 97–106. URL: https://doi.org/10.1007/s10772-009-9039-3.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of l2 comprehensibility and accentedness. *System*, *38*, 301–315. URL: https://doi.org/10.1016/j.system.2010.01.005.

Kendall, T. S. (2009). *Speech rate, pause, and linguistic variation: An examination through the sociolinguistic archive and analysis project*. Ph.D. thesis Duke University. URL: https://hdl.handle.net/10161/1097.

Kienast, M., & Glitza, F. (2003). Respiratory sounds as an idiosyncratic feature in speaker recognition. In *Proc. 15th International Congress of Phonetic Sciences (ICPhS '03)* (pp. 1607–1610). Barcelona. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_1607.html.

Kilbon, J. (2022). *L2 student perceptions regarding their comprehension of academic lectures–a longitudinal study*. Ph.D. thesis University of Leicester. URL: https://doi.org/10.25392/leicester.data.19096304.v1.

King, S. (2015). A reading list of recent advances in speech synthesis. In *Proc. 18th International Congress of Phonetic Sciences (ICPhS '15)*. Glasgow. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1043.pdf.

Kirkland, A., Lameris, H., Székely, É., & Gustafson, J. (2022). Where's the uh, hesitation? The interplay between filled pause location, speech rate and fundamental frequency in perception of confidence. In *Proc. Interspeech 2022* (pp. 4990–4994). URL: https://doi.org/10.21437/Interspeech.2022-10973.

Kjellmer, G. (2003). Hesitation. in defence of er and erm. *English Studies*, *84*, 170–198. URL: https://doi.org/10.1076/enst.84.2.170.14903.

Klatt, D. (1982). The KLATTALK text-to-speech conversion system. In *ICASSP 1982 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1589–1592). URL: https://doi.org/10.1109/ICASSP.1982.1171431.

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, *33*, 17022–17033. URL: https://papers.nips.cc/paper_files/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html.

Kosmala, L. (2020). On the distribution of clicks and inbreaths in class presentations and spontaneous conversations: blending vocal and kinetic activities. In *Proc. Laughter and Other Non-Verbal Vocalisations Workshop ('20)* (pp. 76–79). URL: https://doi.org/10.4119/lw2020-933.

Kosmala, L., & Morgenstern, A. (2019). Should 'uh' and 'um' be categorized as markers of disfluency? the use of fillers in a challenging conversational context. In *Fluency and Disfluency across Languages and Language Varieties*. Presses universitaires de Louvain.

Krikke, T. F., & Truong, K. P. (2013). Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech. In *Proc. Interspeech 2013* (pp. 163–167). URL: https://doi.org/10.21437/Interspeech.2013-59.

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*, 162–179. URL: https://doi.org/10.1016/j.wocn.2006.04.001.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. URL: https://doi.org/10.18637/jss.v082.i13.

Laserna, C. M., Seih, Y.-T., & Pennebaker, J. W. (2014). Um . . . who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, *33*, 328–338. URL: https://doi.org/10.1177/0261927X14526993.

de Leeuw, E. (2007). Hesitation markers in English, German, and Dutch. *Journal of Germanic Linguistics*, *19*, 85–114. URL: https://doi.org/10.1017/S1470542707000049.

Lei, B., Rahman, S. A., & Song, I. (2014). Content-based classification of breath sound with enhanced features. *Neurocomputing*, *141*, 139–147. URL: https://doi.org/10.1016/j.neucom.2014.04.002.

Lenth, R. V. (2023). *Emmeans: Estimated Marginal Means, aka Least-Squares Means*. URL: https://CRAN.R-project.org/package=emmeans r package version 1.8.4-1.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104. URL: https://doi.org/10.1016/0010-0277(83)90026-4.

Lo, J. J. H. (2020). Between äh(m) and euh(m): The distribution and realization of filled pauses in the speech of german-french simultaneous bilinguals. *Language and Speech*, *63*, 746–768. URL: https://doi.org/10.1177/0023830919890068.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in psychology*, *6*, 1171. URL: https://doi.org/10.3389/fpsyg.2015.01171.

Lu, L., Liu, L., Hussain, M. J., & Liu, Y. (2020). I sense you by breath: Speaker recognition via breath biometrics. *IEEE Transactions on Dependable and Secure Computing*, *17*, 306–319. URL: https://doi.org/10.1109/TDSC.2017.2767587.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia*, *48*, 3982–3992. URL: https://doi.org/10.1016/j.neuropsychologia.2010.09.024.

MacIntyre, A. D., & Scott, S. K. (2022). Listeners are sensitive to the speech breathing time series: Evidence from a gap detection task. *Cognition*, *225*, 105171. URL: https://doi.org/10.1016/j.cognition.2022.105171.

Mazzocconi, C., Tian, Y., & Ginzburg, J. (2022). What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*, *13*, 1302–1321. URL: https://doi.org/10.1109/TAFFC.2020.2994533.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (pp. 18–24). URL: https://doi.org/10.25080/Majora-7b98e3ed-003.

McLeod, S. A. (2008). Serial position effect. URL: https://www.simplypsychology.org/primacy-recency.html.

Mendelson, J., & Aylett, M. P. (2017). Beyond the listening test: An interactive approach to tts evaluation. In *Proc. Interspeech 2017* (pp. 249–253). URL: https://doi.org/10.21437/Interspeech.2017-1438.

Merriman, J. (2008). European civilization, 1648-1945 (Yale university: Open Yale courses). https://oyc.yale.edu/history/hist-202. License: Creative Commons BY-NC-SA.

Meurers, D. (2020). Natural language processing and language learning. In *The Concise Encyclopedia of Applied Linguistics* (pp. 817–831). Wiley.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*, 81–97. URL: https://doi.org/10.1037/h0043158.

Mizuno, K. (1990). The effect of pauses on listening. *CASELE Research Bulletin*, *20*, 89–94. URL: https://doi.org/10.18983/casele.20.0_89.

Moniz, H., Batista, F., Mata, A. I., & Trancoso, I. (2014). Speaking style effects in the production of disfluencies. *Speech Communication*, *65*, 20–35. URL: https://doi.org/10.1016/j.specom.2014.05.004.

Moreno, J. (2019). [!] What's the name of it? [!]: Phonetic clicks in word search strategies in Glasgow. In *Proc. 19th International Congress of Phonetic Sciences (ICPhS '19)* (pp. 1823–1827). Melbourne. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/.

Muhlack, B. (2023). Filler particles in english and spanish l1 and l2 speech. In *Proc. 20th International Congress of Phonetic Sciences (ICPhS '23)* (pp. 2418–2422). Prague. URL: https://guarant.cz/icphs2023/110.pdf.

Muhlack, B., Trouvain, J., & Jessen, M. (2023). Distributional and acoustic characteristics of filler particles in german with consideration of forensic-phonetic aspects. *Languages*, *8*. URL: https://doi.org/10.3390/languages8020100.

Nakamura, S., Ishi, C. T., & Kawahara, T. (2020). Analysis and modeling of between-sentence pauses in news speech by japanese newscasters. In *Proc. Speech Prosody 2020* (pp. 680–684). URL: https://doi.org/10.21437/SpeechProsody.2020-139.

Niebuhr, O., & Fischer, K. (2019). Do not hesitate! — unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance. In *Proc. Interspeech 2019* (pp. 544–548). URL: https://doi.org/10.21437/Interspeech.2019-1194.

Ogden, R. (2013). Clicks and percussives in English conversation. *Journal of the International Phonetic Association*, *43*, 299–320. URL: https://doi.org/10.1017/S0025100313000224.

Ogden, R. (2020). Audibly not saying something with clicks. *Research on Language and Social Interaction*, *53*, 66–89. URL: https://doi.org/10.1080/08351813.2020.1712960.

van Os, M., de Jong, N. H., & Bosker, H. R. (2020). Fluency in dialogue: Turn-taking behavior shapes perceived fluency in native and nonnative speech. *Language Learning*, *70*, 1183–1217. URL: https://doi.org/10.1111/lang.12416.

Park, K., & Kim, J. (2019). g2p_en: A simple python module for English grapheme to phoneme conversion. URL: https://github.com/Kyubyong/g2p.

Pinto, D., & Vigil, D. (2019). Searches and clicks in Peninsular Spanish. *Pragmatics*, *29*, 83–106. URL: https://doi.org/10.1075/prag.18020.pin.

Prolific (2014). Prolific. https://www.prolific.co.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Reichel, U. D., Weiss, B., & Michael, T. (2019). Filled pause detection by prosodic discontinuity features. In *Proc. 30th Conference Elektronische Sprachsignalverarbeitung (ESSV '19)* (pp. 272–279). URL: https://www.essv.de/paper.php?id=92.

Rose, R. (2017). Silent and filled pauses and speech planning in first and second language production. In *Proc. Disfluency in Spontaneous Speech (DiSS '17)* (pp. 49–52). URL: https://www.isca-speech.org/archive/pdfs/diss_2017/diss2017_proceedings.pdf.

Rose, R. L. (2013). Crosslinguistic corpus of hesitation phenomena: a corpus for investigating first and second language speech performance. In *Proc. Interspeech 2013* (pp. 992–996). URL: https://doi.org/10.21437/Interspeech.2013-175.

Ruinskiy, D., & Lavner, Y. (2007). An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 838–850. URL: https://doi.org/10.1109/TASL.2006.889750.

Saraiva, A. A., Santos, D., Francisco, A., Sousa, J. V. M., Ferreira, N. M. F., Soares, S., & Valente, A. (2020). Classification of respiratory sounds with convolutional neural network. In *Proc. 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - BIOINFORMATICS* (pp. 138–144). URL: https://doi.org/10.5220/0008965101380144.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*, 1141–1152. URL: https://doi.org/10.1111/2041-210X.13434.

Schröder, M., & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, *6*, 365–377. URL: https://doi.org/10.1023/A:1025708916924.

Schulte, M. (2020). Functions and social meanings of click sounds in irish english. In *Proc. Laughter and Other Non-Verbal Vocalisations Workshop ('20)* (pp. 32–35). URL: https://doi.org/10.4119/lw2020-921.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y.

(2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4779–4783). URL: https://doi.org/10.1109/ICASSP.2018.8461368.

Silber-Varod, V., Amit, D., & Lerner, A. (2020). Tracing changes over the course of the conversation: A case study on filled pauses rates. In *Proc. Speech Prosody 2020* (pp. 754–758). URL: https://doi.org/10.21437/SpeechProsody.2020-154.

Singh, L. G., Adiga, N., Sharma, B., Singh, S. R., & Prasanna, S. (2017). Automatic pause marking for speech synthesis. In *TENCON 2017 IEEE Region 10 Conference* (pp. 1790–1794). URL: https://doi.org/10.1109/TENCON.2017.8228148.

Székely, É., Henter, G. E., Beskow, J., & Gustafson, J. (2019a). How to train your fillers: uh and um in spontaneous speech synthesis. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)* (pp. 245–250). URL: https://doi.org/10.21437/SSW.2019-44.

Székely, É., Henter, G. E., Beskow, J., & Gustafson, J. (2020). Breathing and speech planning in spontaneous speech synthesis. In *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7649–7653). URL: https://doi.org/10.1109/ICASSP40776.2020.9054107.

Székely, É., Henter, G. E., & Gustafson, J. (2019b). Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector. In *ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6925–6929). URL: https://doi.org/10.1109/ICASSP.2019.8683846.

Székely, É., Mendelson, J., & Gustafson, J. (2017). Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies. In *Proc. Interspeech 2017* (pp. 804–808). URL: https://doi.org/10.21437/Interspeech.2017-1507.

Taylor, P., Black, A. W., & Caley, R. (1998). The architecture of the Festival speech synthesis system. In *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis (SSW 3)* (pp. 147–152). URL: https://www.isca-speech.org/archive/ssw_1998/taylor98_ssw.html.

Torchiano, M. (2020). *effsize: Efficient Effect Size Computation*. URL: https://CRAN.R-project.org/package=effsize. doi:doi: 10.5281/zenodo.1480624 r package version 0.8.1.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of l2 experience on prosody and fluency characteristics of l2 speech. *Studies in Second Language Acquisition*, *28*, 1–30. URL: https://doi.org/10.1017/S0272263106060013.

Trouvain, J. (2011). Between excitement and triumph-live football commentaries in radio vs. tv. In *Proc. 17th International Congress of Phonetic Sciences (ICPhS '11)* (pp. 2022–2025). Hong Kong. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Trouvain/Trouvain.pdf.

Trouvain, J., & Barry, W. J. (2000). The prosody of excitement in horse race commentaries. In *Proc. ITRW on Speech and Emotion* (pp. 86–91). URL: https://www.isca-speech.org/archive/speechemotion_2000/trouvain00_speechemotion.html.

Trouvain, J., & Malisz, Z. (2016). Inter-speech clicks in an Interspeech keynote. In *Proc. Interspeech 2016* (pp. 1397–1401). URL: https://doi.org/10.21437/Interspeech.2016-1064.

Trouvain, J., & Möbius, B. (2013). Einatmungsgeräusche vor synthetisch erzeugten Sätzen: Eine Pilotstudie. In *Proc. 24th Conference Elektronische Sprachsignalverarbeitung (ESSV '13)* (pp. 50–55). URL: https://www.essv.de/paper.php?id=105.

Trouvain, J., & Möbius, B. (2018). Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache. In *Proc. 29th Conference Elektronische Sprachsignalverarbeitung (ESSV '18)* (pp. 334–341). URL: https://www.essv.de/paper.php?id=426.

Trouvain, J., & Truong, K. P. (2012). Comparing non-verbal vocalisations in conversational speech corpora. In *Proc. of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals* (pp. 36–39). URL: http://www.lrec-conf.org/proceedings/lrec2012/workshops/18.ProceedingsES32012.pdf.

Trouvain, J., & Werner, R. (2022). A phonetic view on annotating speech pauses and pause-internal phonetic particles. *Transkription und Annotation Gesprochener Sprache und Multimodaler Interaktion: Konzepte, Probleme, Lösungen*, *64*, 55–73.

Trouvain, J., Werner, R., & Möbius, B. (2020). An acoustic analysis of inbreath noises in read and spontaneous speech. In *Proc. Speech Prosody 2020* (pp. 789–793). URL: https://doi.org/10.21437/SpeechProsody.2020-161.

Vigil, D., & Pinto, D. (2020). An experimental study of the detection of clicks in English. *Pragmatics & Cognition*, *27*, 457–473. URL: https://doi.org/10.1075/pc.20009.vig.

Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, *22*, 129–144. URL: https://doi.org/10.1177/002383097902200203.

Wagner, P., & Betz, S. (2017). Speech synthesis evaluation: Realizing a social turn. In *Proc. 28th Conference Elektronische Sprachsignalverarbeitung (ESSV '17)* (pp. 167–173). URL: https://www.essv.de/paper.php?id=234.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12. URL: https://doi.org/10.1016/j.wocn.2014.11.001.

Wang, S., Alexanderson, S., Gustafson, J., Beskow, J., Henter, G. E., & Székely, E. (2021). Integrated speech and gesture synthesis. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)* (pp. 177–185). URL: https://doi.org/10.1145/3462244.3479914.

Wang, S., Gustafson, J., & Székely, É. (2022). Evaluating sampling-based filler insertion with spontaneous tts. In *Proc. 13th Conference on Language Resources and Evaluation (LREC '22)* (pp. 1960–1969). Marseille. URL: https://aclanthology.org/2022.lrec-1.210.

Wang, Y.-T., Green, J. R., Nip, I. S., Kent, R. D., & Kent, J. F. (2010). Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatrica et Logopaedica*, *62*, 297–302. URL: https://doi.org/10.1159/000316976.

Wargo, J. (2010). Environmental politics and law (Yale university: Open Yale courses). https://oyc.yale.edu/environmental-studies/evst-255. License: Creative Commons BY-NC-SA.

Watanabe, M., Den, Y., Hirose, K., & Minematsu, N. (2005). The effects of filled pauses on native and non-native listeners' speech processing. In *Proc. Disfluency in Spontaneous Speech (DiSS '05)* (pp. 169–172). URL: https://www.isca-speech.org/archive/diss_2005/watanabe05_diss.html.

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*, 81–94. URL: https://doi.org/10.1016/j.specom.2007.06.002.

Watanabe, M., Hirose, K., Den, Y., Miwa, S., & Minematsu, N. (2006). Factors influencing ratios of filled pauses at clause boundaries in Japanese. In *Proc. First ITRW on Experimental Linguistics* (pp. 253–256). URL: https://www.isca-speech.org/archive/exling_2006/watanabe06_exling.html.

Werner, R., Fuchs, S., Trouvain, J., & Möbius, B. (2021). Inhalations in speech: Acoustic and physiological characteristics. In *Proc. Interspeech 2021* (pp. 3186–3190). URL: https://doi.org/10.21437/Interspeech.2021-1262.

Werner, R., Trouvain, J., & Möbius, B. (2022). Optionality and variability of speech pauses in read speech across languages and rates. In *Proc. Speech Prosody 2022* (pp. 312–316). URL: https://doi.org/10.21437/SpeechProsody.2022-64.

Whalen, D., & Kinsella-Shaw, J. (1997). Exploring the relationship of inspiration duration to utterance duration. *Phonetica*, *54*, 138–152. URL: https://doi.org/10.1159/000262218.

Whalen, D. H., Hoequist, C. E., & Sheffert, S. M. (1995). The effects of breath sounds on the perception of synthetic speech. *The Journal of the Acoustical Society of America*, *97*, 3147–3153. URL: https://doi.org/10.1121/1.411875.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org.

Wickham, H. (2022). *stringr: Simple, Consistent Wrappers for Common String Operations*. URL: https://CRAN.R-project.org/package=stringr r package version 1.5.0.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023a). *dplyr: A Grammar of Data Manipulation*. URL: https://CRAN.R-project.org/package=dplyr r package version 1.1.1.

Wickham, H., Vaughan, D., & Girlich, M. (2023b). *tidyr: Tidy Messy Data*. URL: https://CRAN.R-project.org/package=tidyr r package version 1.3.0.

Winter, B., & Grawunder, S. (2012). The phonetic profile of Korean formal and informal speech registers. *Journal of Phonetics*, *40*, 808–815. URL: https://doi.org/10.1016/j.wocn.2012.08.006.

Włodarczak, M., Heldner, M., & Edlund, J. (2015). Communicative needs and respiratory constraints. In *Proc. Interspeech 2015* (pp. 3051–3055). URL: https://doi.org/10.21437/Interspeech.2015-620.

Wrightson, K. (2009). Early modern England: Politics, religion, and society under the Tudors and Stuarts (Yale university: Open Yale courses). https://oyc.yale.edu/history/hist-251. License: Creative Commons BY-NC-SA.

Yang, X., Xu, M., & Yang, Y. (2014). Predictors of pause duration in read-aloud discourse. *IEICE Transactions on Information and Systems*, *E97.D*, 1461–1467. URL: https://doi.org/10.1587/transinf.E97.D.1461.

Zayats, V., Tran, T., Wright, R., Mansfield, C., & Ostendorf, M. (2019). Disfluencies and human speech transcription errors. In *Proc. Interspeech 2019* (pp. 3088–3092). URL: https://doi.org/10.21437/Interspeech.2019-3134.

Zellers, M. (2022). An overview of discourse clicks in central Swedish. In *Proc. Interspeech 2022* (pp. 3423–3427). URL: https://doi.org/10.21437/Interspeech.2022-583.