NATURAL LANGUAGE PROCESSING

A Literature Review

Rahul S. Agasthya

Stony Brook University, New York.

ABSTRACT

Ever since the publication of Alan Turning's *Computing Machinery and Intelligence*, research and development has steadily shifted towards the field of Natural Language Processing. Text to Speech Conversion and Text Recognition have been a major area of interest. However, it was not only until the mid-1980s that the first statistical machine translation systems were developed. The most notable success in the field of Speech Processing and Machine Learning was the IBM's Watson, which successfully understood questions and produced correct answers, and converted to Speech. All the Speech Processing software have limitations like understanding human language, deciphering punctuations and sentence structuring. Most speech recognition researchers who understood such barriers moved away from neural nets to pursue generative modeling approaches until the recent resurgence of deep learning in 2009. Research scholars Geoffrey Hinton and Deng Min, from the University of Toronto in collaboration with University of Toronto, Microsoft, Google, and IBM, ignited a renaissance of applications of deep feedforward neural networks to speech recognition. Text to speech conversion and speech recognition, language translation, etc. have become popular NLP software. This paper talks about the working of Test to Speech Conversion and Speech Recognition software and modern tools that use NLP.

Keywords: Language; Speech Processing; Natural Language Processing; Machine Learning; Computer Understanding; Part of Speech Tagging; Syllable Tagging.

TABLE OF CONTENTS

WHAT IS NLP?

Natural Language Processing (NLP) is the science most directly associated to processing human (natural) language. It derives from computer science since computers, or any other processing unit, are the target devices used to accomplish such processing. This description responds basically to the "Processing" particle in NLP. What makes NLP different from any other processing-related activity is the field of application: the human languages. They deal with more knowledge-related aspects thus requiring the support of learning capabilities by the processors of text.

The following is a list of some of the most commonly researched tasks in NLP. Note that some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks.

Though NLP tasks are obviously very closely intertwined, they are frequently, for convenience, subdivided into categories. A coarse division is given below:

SYNTAX

LEMMATIZATION

In linguistics, Lemmatization is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form. In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research. (Green, Breimyer, Kumar, & Samatova)

Morphological Segmentation.

In linguistics, morphology is the study of words, how they are formed, and their relationship to other words in the same language. It analyzes the structure of words and parts of words, such as stems, root words, prefixes, and suffixes. Morphology also looks at parts of speech, intonation and stress, and the ways context can change a word's pronunciation and meaning. Morphology differs from morphological typology, which is the classification of languages based on their use of words and lexicology, which is the study of words and how they make up a language's vocabulary. (Anderson S. R., 2016)

Procedure in NLP

Separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology (i.e. the structure of words) of the language being considered. English has fairly simple morphology, especially inflectional morphology, and thus it is often possible to ignore this task entirely and simply model all possible forms of a word (e.g. "open, opens, opened, opening") as separate words. In languages such as Turkish or Meitei, a highly agglutinated Indian language, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms. (Kishorjit, Rai, & Sivaji, 2012)

Part of Speech Tagging.

In corpus linguistics, part-of-speech tagging, also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is

commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Procedure in NLP

Given a sentence, determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, “book” can be a noun (“the book on the table”) or verb (“to book a flight”); “set” can be a noun, verb or adjective; and “out” can be any of at least five different parts of speech. Some languages have more such ambiguity than others. Languages with little inflectional morphology, such as English are particularly prone to such ambiguity. Chinese is prone to such ambiguity because it is a tonal language during verbalization. Such inflection is not readily conveyed via the entities employed within the orthography to convey intended meaning. (Church, 1988)

Parsing.

Parsing is the process of analyzing a string of symbols, either in natural language or in computer languages, conforming to the rules of a formal grammar. Within computational linguistics the term is used to refer to the formal analysis by a computer of a sentence or other string of words into its constituents, resulting in a parse tree showing their syntactic relation to each other, which may also contain semantic and other information.

Procedure in NLP

Determine the parse tree (grammatical analysis) of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. In fact, perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human).

Sentence Breaking.

Sentence Breaking is the problem in NLP of deciding where sentences begin and end. Often natural language processing tools require their input to be divided into sentences for a number of reasons. However, sentence boundary identification is challenging because punctuation marks are often ambiguous. For example, a period may denote an abbreviation, decimal point, an ellipsis, or an email address – not the end of a sentence. About 47% of the periods in the Wall Street Journal corpus denote abbreviations. As well, question marks and exclamation marks may appear in embedded quotations, emoticons, computer code, and slang. (Stamatatos, Fakotakis, & Kokkinakis, 2009).

Procedure in NLP

Given a chunk of text, find the sentence boundaries. Sentence boundaries are often marked by periods or other punctuation marks, but these same characters can serve other purposes (e.g. marking abbreviations).

STEMMING

In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation. Stemming programs are commonly referred to as stemming algorithms or stemmers. (Lovins, 1968)

Word Segmentation.

Word segmentation is the problem of dividing a string of written language into its component words. In English and many other languages using some form of the Latin alphabet, the space is a good approximation of a word divider (word delimiter). However, the equivalent to this character is not found in all written scripts, and without it word segmentation is a difficult problem. Languages which do not have a trivial word segmentation process include Chinese, Japanese, where sentences but not words are delimited, Thai and Lao, where phrases and sentences but not words are delimited, and Vietnamese, where syllables but not words are delimited. In some writing systems however, such as the Ge'ez script used for Amharic and Tigrinya among other languages, words are explicitly delimited with a non-whitespace character.

The Unicode Consortium has published a Standard Annex on Text Segmentation, exploring the issues of segmentation in multiscript texts. Word splitting is the process of parsing concatenated text (i.e. text that contains no spaces or other word separators) to infer where word breaks exist. Word splitting may also refer to the process of hyphenation. (Freddy, 2000)

Procedure in NLP

Separate a chunk of continuous text into separate words. For a language like English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language.

Terminology Extraction.

Terminology extraction is a subtask of information extraction. The goal of terminology extraction is to automatically extract relevant terms from a given corpus.

In the semantic web era, a growing number of communities and networked enterprises started to access and interoperate through the internet. Modeling these communities and their information needs is important for several web applications, like topic-driven web crawlers, web services, recommender systems, etc. The development of terminology extraction is essential to the language industry. (Park, Byrd, & Boguraev, 2002)

Procedure in NLP

The goal of terminology extraction is to automatically extract relevant terms from a given corpus.

SEMANTICS

LEXICAL SEMANTICS

Lexical Semantics, is a subfield of linguistic semantics. The units of analysis in lexical semantics are lexical units which include not only words but also sub-words or sub-units such as affixes and even compound words and phrases. Lexical units make up the catalogue of words in a language, the lexicon. Lexical semantics looks at how the meaning of the lexical units correlates with the structure of the language or syntax. This is referred to as syntax-semantic interface. (Pustejovsky, 1995)

Procedure in NLP

Generate the computational meaning of individual words in context.

MACHINE TRANSLATION

Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

On a basic level, Machine translation performs simple substitution of words in one language for words in another, but that alone usually cannot produce a good translation of a text because recognition of whole phrases and their closest counterparts in the target language is needed.

Solving this problem with corpus statistical, and neural techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies. (Albat, 2012)

Procedure in NLP

Automatically translate text from one human language to another. This is one of the most difficult problems, and is a member of a class of problems colloquially termed "AI-complete," i.e. requiring all of the different types of knowledge that humans possess (grammar, semantics, facts about the real world, etc.) in order to solve properly.

NAMED ENTITY RECOGNITION

Named-entity recognition (NER), also known as entity identification, entity chunking and entity extraction is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Most research on NER systems has been structured as taking an unannotated block of text, such as this one:

Jim bought 300 shares of Acme Corp. in 2006.

And producing an annotated block of text that highlights the names of entities:

[Jim] PERSON bought 300 shares of [Acme Corp.] ORGANIZATION in [2006] TIME.

In this example, a person name consisting of one token, a two-token company name and a temporal expression have been detected and classified. State-of-the-art Named-entity recognition systems for English produce near-human performance. For example, the best system entering MUC-7 scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%. (Elaine & Perzanowski, 1998)

Procedure in NLP

Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization). Note that, although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case, is often inaccurate or insufficient. For example, the first word of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized. Furthermore, many other languages in non-Western scripts (e.g. Chinese or Arabic) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names. For example, German capitalizes all nouns, regardless of whether they refer to names, and French and Spanish do not capitalize names that serve as adjectives.

NATURAL LANGUAGE GENERATION

Natural language generation (NLG) is the natural language processing task of generating natural language from a machine representation system such as a knowledge base or a logical form. Psycholinguists prefer the term language production when such formal representations are interpreted as models for mental representations. NLG may be viewed as the opposite of natural language understanding: whereas in natural language understanding the system needs to disambiguate the input sentence to produce the machine representation language, in NLG the system needs to make decisions about how to put a concept into words. (R & P)

Procedure in NLP

Convert information from computer databases or semantic intents into readable human language.

NATURAL LANGUAGE UNDERSTANDING

Natural language understanding (NLU) is a subtopic of natural language processing in artificial intelligence that deals with machine reading comprehension. NLU is considered an AI-hard problem. The process of disassembling and parsing input is more complex than the reverse process of assembling output in natural language generation because of the occurrence of unknown and unexpected features in the input and the need to determine the appropriate syntactic and semantic schemes to apply to it, factors which are pre-determined when outputting language. There is considerable commercial interest in the field because of its application to news-gathering, text categorization, voice-activation, archiving, and large-scale content-analysis. (Babrow)

Procedure in NLP

Convert chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate. Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural languages concepts. Introduction and creation of language metamodel and ontology are efficient however empirical solutions. An explicit formalization of natural languages semantics without confusions with implicit assumptions such as closed-world assumption (CWA) vs. open-world assumption, or subjective Yes/No vs. objective True/False is expected for the construction of a basis of semantics formalization.

OPTICAL CHARACTER RECOGNITION

Optical character recognition is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo)

or from subtitle text superimposed on an image (for example from a television broadcast). It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. (Schantz, 2012)

Procedure in NLP

Given an image representing printed text, determine the corresponding text.

QUESTION ANSWERING

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language. A QA implementation, usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. More commonly, QA systems can pull answers from an unstructured collection of natural language documents. Some examples of natural language document collections used for QA systems include: (Morante, Krallinger, Valencia, & Daelemans, CLEF)

1. A local collection of reference texts
2. Internal organization documents and web pages
3. Compiled newswire reports
4. A subset of World Wide Web pages

Procedure in NLP

Given a human-language question, determine its answer. Typical questions have a specific right answer (such as “What is the capital of Canada?”), but sometimes open-ended questions are also considered (such as “What is the meaning of life?”). Recent works have looked at even more complex questions.

RECOGNIZING TEXTUAL ENTAILMENT

Textual Entailment in natural language processing is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text. In the Textual Entailment framework, the entailing and entailed texts are termed text (t) and hypothesis (h), respectively. Textual entailment is not the same as pure logical entailment- it has a more relaxed definition: “t entails h” ($t \Rightarrow h$) if, typically, a human reading t would infer that h is most likely true. The relation is directional because even if “t entails h,” the reverse “h entails t” is much less certain. (Dagan, Clickman, & Magnini, 2004)

Procedure in NLP.

Given two text fragments, determine if one being true entails the other, entails the other’s negation, or allows the other to be either true or false.

RELATIONSHIP EXTRACTION

A relationship extraction task requires the detection and classification of semantic relationship mentions within a set of artifacts, typically from text or XML documents. The task is very similar to that of Information Extraction (IE), but IE additionally requires the removal of repeated relations and generally refers to the extraction of many different relationships. Application domains where relationship extraction is useful include gene-disease relationships, protein-protein interaction, etc. (Chun, et al., 2006)

Procedure in NLP

Given a chunk of text, identify the relationships among named entities (e.g. who is married to whom).

SENTIMENT ANALYSIS

Sentiment analysis (sometimes known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. (Stone, Dunphy, & Smith, 1966)

Procedure in NLP

Extract subjective information usually from a set of documents, often using online reviews to determine “polarity” about specific objects. It is especially useful for identifying trends of public opinion in the social media, for the purpose of marketing.

TOPIC SEGMENTATION AND RECOGNITION

Text segmentation is the process of dividing written text into meaningful units, such as words, sentences, or topics. The term applies both to mental processes used by humans when reading text, and to artificial processes implemented in computers, which are the subject of natural language processing. The problem is non-trivial, because while some written languages have explicit word boundary markers, such as the word spaces of written English and the distinctive initial, medial and final letter shapes of Arabic, such signals are sometimes ambiguous and not present in all written languages. (Freddy, 2000)

Procedure in NLP

Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.

WORD SENSE DISAMBIGUATION

In computational linguistics, word-sense disambiguation (WSD) is an open problem of natural language processing and ontology. Word-sense disambiguation is identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings. The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, anaphora resolution, coherence, inference, etcetera.

The human brain is quite proficient at word-sense disambiguation. The fact that natural language is formed in a way that requires so much of it is a reflection of that neurologic reality. In other words, human language developed in a way that reflects (and also has helped to shape) the innate ability provided by the brain's neural networks. In computer science and the information technology that it enables, it has been a long-term challenge to develop the ability in computers to do natural language processing and machine learning.

To date, a rich variety of techniques have been researched, from dictionary-based methods that use the knowledge encoded in lexical resources, to supervised machine learning methods in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely unsupervised methods that cluster occurrences of words, thereby inducing word senses. Among these, supervised learning approaches have been the most successful algorithms to date.

Accuracy of current algorithms is difficult to state without a host of caveats. In English, accuracy at the coarse-grained (homograph) level is routinely above 90%, with some methods on

particular homographs achieving over 96%. On finer-grained sense distinctions, top accuracies from 59.1% to 69.0% have been reported in recent evaluation exercises (SemEval-2007, Senseval-2), where the baseline accuracy of the simplest possible algorithm of always choosing the most frequent sense was 51.4% and 57%, respectively. (Moro, Raganato, & Navigli, 2014)

Procedure in NLP

Many words have more than one meaning; we have to select the meaning which makes the most sense in context. For this problem, we are typically given a list of words and associated word senses, e.g. from a dictionary or from an online resource such as WordNet.

DISCOURSE

AUTOMATIC SUMMARIZATION

Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. (Camargo & Gonzalez, 2009)

Procedure in NLP

Produce a readable summary of a chunk of text. Often used to provide summaries of text of a known type, such as articles in the financial section of a newspaper.

COREFERENCE RESOLUTION

In linguistics, coreference, sometimes written co-reference, occurs when two or more expressions in a text refer to the same person or thing; they have the same referent, e.g. *Bill* said *he* would come; the proper noun *Bill* and the pronoun *he* refers to the same person, namely to *Bill*. Coreference is the main concept underlying binding phenomena in the field of syntax. The theory of binding explores the syntactic relationship that exists between coreferential expressions in

sentences and texts. When two expressions are coreferential, the one is usually a full form (the antecedent) and the other is an abbreviated form (a proform or anaphor). Linguists use indices to show coreference, as with the *i* index in the example $Bill_i$ said he_i would come. The two expressions with the same reference are *coindexed*, hence in this example *Bill* and *he* are coindexed, indicating that they should be interpreted as coreferential.

Procedure in NLP

Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities"). Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names to which they refer. The more general task of coreference resolution also includes identifying so-called "bridging relationships" involving referring expressions. For example, in a sentence such as "He entered John's house through the front door," "the front door" is a referring expression and the bridging relationship to be identified is the fact that the door being referred to is the front door of John's house (rather than of some other structure that might also be referred to).

DISCOURSE ANALYSIS

Discourse analysis (DA), or discourse studies, is a general term for a number of approaches to analyze written, vocal, or sign language use, or any significant semiotic event. The objects of discourse analysis (discourse, writing, conversation, communicative event) are variously defined in terms of coherent sequences of sentences, propositions, speech, or turns-at-talk. Contrary to much of traditional linguistics, discourse analysts not only study language use 'beyond the sentence boundary' but also prefer to analyze 'naturally occurring' language use, not invented examples. (Discourse Analysis - What Speakers Do in Conversion , 2016)

Procedure in NLP

This rubric includes a number of related tasks. One task is identifying the discourse structure of connected text, i.e. the nature of the discourse relationships between sentences (e.g. elaboration, explanation, contrast). Another possible task is recognizing and classifying the speech acts in a chunk of text (e.g. yes-no question, content question, statement, assertion, etc.).

NLP IN SPEECH

Natural Language Processing and Spoken Language Processing involves computational approaches to the analysis and generation of text and speech. At Columbia, these include text and speech summarization, question answering, machine translation, syntax and parsing, language generation, spoken dialogue systems, semantic representation and analysis, and the study of emotional and deceptive speech, in English, Arabic, and Mandarin, inter alia.

SPEECH RECOGNITION

Speech recognition (SR) is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as "automatic speech recognition" (ASR), "computer speech recognition," or just "speech to text" (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields.

Some SR systems use "training" (also called "enrollment") where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent." (FGC/SRS Speech Recognition Server, 2007)

Speech recognition applications include voice user interfaces such as voice dialing (e.g. "Call home"), call routing (e.g. "I would like to make a collect call"), domotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed Direct Voice Input).

The term voice recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process. (Dictionary, 2012)

From the technology perspective, speech recognition has a long history with several waves of major innovations. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidenced not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems. These speech industry players include Google, Microsoft, IBM, Baidu, Apple, Amazon, Nuance, SoundHound, IflyTek, CDAC many of which have publicized the core technology in their speech recognition systems as being based on deep learning.

SPEECH SEGMENTATION

Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. The term applies both to the mental processes used by humans, and to artificial processes of natural language processing.

Speech segmentation is a subfield of general speech perception and an important subproblem of the technologically focused field of speech recognition, and cannot be adequately solved in isolation. As in most natural language processing problems, one must take into account context, grammar, and semantics, and even so the result is often a probabilistic division (statistically based on likelihood) rather than a categorical one. Though it seems that coarticulation— a phenomenon which may happen between adjacent words just as easily as within

a single word— presents the main challenge in speech segmentation across languages, some other problems and strategies employed in solving those problems can be seen in the following sections.

Lexical Recognition

In natural languages, the meaning of a complex spoken sentence can be understood by decomposing it into smaller lexical segments (roughly, the words of the language), associating a meaning to each segment, and combining those meanings according to the grammar rules of the language.

Though lexical recognition is not thought to be used by infants in their first year, due to their highly limited vocabularies, it is one of the major processes involved in speech segmentation for adults. Three main models of lexical recognition exist in current research: first, whole-word access, which argues that words have a whole-word representation in the lexicon; second, decomposition, which argues that morphologically complex words are broken down into their morphemes (roots, stems, inflections, etc.) and then interpreted and; third, the view that whole-word and decomposition models are both used, but that the whole-word model provides some computational advantages and is therefore dominant in lexical recognition. (Badecker & Allen, 2002)

To give an example, in a whole-word model, the word “cats” might be stored and searched for by letter, first “c,” then “ca,” “cat,” and finally “cats.” The same word, in a decomposition model, would likely be stored under the root word “cat” and could be searched for after removing the “s” suffix. “Falling,” similarly, would be stored as “fall” and suffixed with the “ing” inflection.

Though proponents of the decomposition model recognize that a morpheme-by-morpheme analysis may require significantly more computation, they argue that the unpacking of

morphological information is necessary for other processes (such as syntactic structure) which may occur parallel to lexical searches.

As a whole, research into systems of human lexical recognition is limited due to little experimental evidence that fully discriminates between the three main models. (Badecker & Allen, 2002)

In any case, lexical recognition likely contributes significantly to speech segmentation through the contextual clues it provides, given that it is a heavily probabilistic system—based on the statistical likelihood of certain words or constituents occurring together. For example, one can imagine a situation where a person might say "I bought my dog at a ____ shop" and the missing word's vowel is pronounced as in "net," "sweat," or "pet." While the probability of "netshop" is extremely low, since "netshop" isn't currently a compound or phrase in English, and "sweatshop" also seems contextually improbable, "pet shop" is a good fit because it is a common phrase and is also related to the word "dog." (Lieberman, Faaborg, Daher, & Espinosa, 2005)

PHONO TACTIC CUES

For most spoken languages, the boundaries between lexical units are difficult to identify; phonotactics are one answer to this issue. One might expect that the inter-word spaces used by many written languages like English or Spanish would correspond to pauses in their spoken version, but that is true only in very slow speech, when the speaker deliberately inserts those pauses. In normal speech, one typically finds many consecutive words being said with no pauses between them, and often the final sounds of one-word blend smoothly or fuse with the initial sounds of the next word.

The notion that speech is produced like writing, as a sequence of distinct vowels and consonants, may be a relic of alphabetic heritage for some language communities. In fact, the way

vowels are produced depends on the surrounding consonants just as consonants are affected by surrounding vowels; this is called coarticulation. For example, in the word “kit,” the [k] is farther forward than when we say ‘caught.’ But also, the vowel in “kick” is phonetically different from the vowel in “kit,” though we normally do not hear this. In addition, there are language-specific changes which occur in casual speech which makes it quite different from spelling. For example, in English, the phrase “hit you” could often be more appropriately spelled “hitcha.”

Vowel harmony in languages like Finnish can also serve to provide phonotactic cues. While the system does not allow front vowels and back vowels to exist together within one morpheme, compounds allow two morphemes to maintain their own vowel harmony while coexisting in a word. Therefore, in compounds such as “selkä/ongelma” (‘back problem’) where vowel harmony is distinct between two constituents in a compound, the boundary will be wherever the switch in harmony takes place—between the “ä” and the “ö” in this case. Still, there are instances where phonotactics may not aid in segmentation. Words with unclear clusters or uncontrasted vowel harmony as in “opinto/uudistus” (‘student reform’) do not offer phonotactic clues as to how they are segmented. (Adriaans & Kager, 2010)

SPEECH SEGMENTATION IN INFANTS AND NON-NATIVES

Infants are one major focus of research in speech segmentation. Since infants have not yet acquired a lexicon capable of providing extensive contextual clues or probability-based word searches within their first year, as mentioned above, they must often rely primarily upon phonotactic and rhythmic cues (with prosody being the dominant cue), all of which are language-specific. Between 6 and 9 months, infants begin to lose the ability to discriminate between sounds not present in their native language and grow sensitive to the sound structure of their native language, with the word segmentation abilities appearing around 7.5 months.

Though much more research needs to be done on the exact processes that infants use to begin speech segmentation, current and past studies suggest that English-native infants approach stressed syllables as the beginning of words. At 7.5 months, infants appear to be able to segment bi-syllabic words with strong-weak stress patterns, though weak-strong stress patterns are often misinterpreted, e.g. interpreting "guiTAR is" as "GUI TARis." It seems that infants also show some complexity in tracking frequency and probability of words, for instance, recognizing that although the syllables "the" and "dog" occur together frequently, "the" also commonly occurs with other syllables, which may lead to the analysis that "dog" is an individual word or concept instead of the interpretation "thedog." (Juscyk & Houston, 1999)

TEXT TO SPEECH

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. (Allen, Hunnicutt, & Klatt, 1987)

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output. (Rubin, Baer, & Mermelstein, 1981)

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood clearly. An intelligible text-to-speech program allows people with

visual impairments or reading disabilities to listen to written words on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s.

PROCESS

A text-to-speech system (or "engine") is composed of two parts (Santen, Sproat, Olive, & Hirschberg, 1997): a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end—often referred to as the synthesizer—then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech. (Santen, Sproat, Olive, & Hirschberg, 1997)

SYNTHESIZER TECHNOLOGIES

The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. The ideal speech synthesizer is both natural and intelligible. Speech synthesis systems usually try to maximize both characteristics.

The two primary technologies generating synthetic speech waveforms are concatenative synthesis and formant synthesis. Each technology has strengths and weaknesses, and the intended uses of a synthesis system will typically determine which approach is used. (Taylor, 2009)

Concatenation Synthesis

Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. There are three main sub-types of concatenative synthesis.

Unit Selection Synthesis

Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some manual correction afterward, using visual representations such as the waveform and spectrogram. (Black, 2002)

Diphone Synthesis

Diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones, and German about 2500. In diphone synthesis, only one example of each diphone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as linear predictive coding, PSOLA or MBROLA. (Dutoit, Pagel, Pierret, Bataille, & Vreckren, 1996)

Domain-specific Synthesis

Domain-specific synthesis concatenates prerecorded words and phrases to create complete utterances. It is used in applications where the variety of texts the system will output is limited to a particular domain, like transit schedule announcements or weather reports. (Lamel, Gauvain, Prouts, Bouchier, & Boesch, 1993)

Formant Synthesis

Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model (Dartmouth College, 2011). Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis; however, many concatenative systems also have rules-based components. Many systems based on formant synthesis technology generate artificial, robotic-sounding speech that would never be mistaken for human speech. However, maximum naturalness is not always the goal of a speech synthesis system, and formant synthesis systems have advantages over concatenative systems. Formant-synthesized speech can be reliably intelligible, even at very high speeds, avoiding the acoustic glitches that commonly plague concatenative systems. High-speed synthesized speech is used by the visually impaired to quickly navigate computers using a screen reader. Formant synthesizers are usually smaller programs than concatenative systems because they do not have a database of speech samples. They can therefore be used in embedded systems, where memory and microprocessor power are especially limited. Because formant-based systems have complete control of all aspects of the output speech, a wide variety of prosodies and intonations can be output, conveying not just questions and statements, but a variety of emotions and tones of voice. (Dartmouth College, 1993)

Articulatory Synthesis

Articulatory synthesis refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there. The first articulatory synthesizer regularly used for laboratory experiments was developed at Haskins Laboratories in the mid-1970s by Philip Rubin, Tom Baer, and Paul Mermelstein. This synthesizer, known as ASY, was based on vocal tract models developed at Bell Laboratories in the 1960s and 1970s by Paul Mermelstein, Cecil Coker, and colleagues.

More recent synthesizers, developed by Jorge C. Lucero and colleagues, incorporate models of vocal fold biomechanics, glottal aerodynamics and acoustic wave propagation in the bronqui, traquea, nasal and oral cavities, and thus constitute full systems of physics-based speech simulation. (Englert, Madazio, Gielow, Lucero, & Behlau, Perceptual error identification of human and synthesized voices, 2015)

HMM-based Synthesis

HMM-based synthesis is a synthesis method based on hidden Markov models, also called Statistical Parametric Synthesis. In this system, the frequency spectrum (vocal tract), fundamental frequency (voice source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion. (Oura, 2017)

Sinewave Synthesis

Sinewave synthesis is a technique for synthesizing speech by replacing the formants (main bands of energy) with pure tone whistles. (Remez, Rubin, Pisoni, & Carrell, 1981)

DEDICATED HARDWARE

Technologies which were once used include, —

1. Icophone
2. Votrax
    (a) SC-01A
    (b) SC-02
3. General Instrument SP0256-AL2
4. National Semiconductor DT1050 Digitalker
5. Silicon Systems SSI 263
6. Texas Instruments LPC Speech Chips
    (a) TMS5110A
    (b) TMS5200
    (c) MSP50C6XX
7. Hitachi HD38880BP

The Current Technologies include, —

1. Magnevation SpeakJet TTS256 Hobby and experimenter
2. Epson S1V30120F01A100
3. Textspeak TTS-EM

HARDWARE AND SOFTWARE SYSTEMS

Popular systems offering speech synthesis as a built-in capability.

Mattel

The Mattel Intellivision game console offered the Intellivoice Voice Synthesis module in 1982. It included the SP0256 Narrator speech synthesizer chip on a removable cartridge. The Narrator had 2kB of Read-Only Memory (ROM), and this was utilized to store a database of generic words that could be combined to make phrases in Intellivision games. Since the Orator

chip could also accept speech data from external memory, any additional words or phrases needed could be stored inside the cartridge itself. The data consisted of strings of analog-filter coefficients to modify the behavior of the chip's synthetic vocal-tract model, rather than simple digitized samples (Milligan, 2016).

SAM

Also released in 1982, Software Automatic Mouth was the first commercial all-software voice synthesis program. It was later used as the basis for Macintalk. The program was available for non-Macintosh Apple computers (including the Apple II, and the Lisa), various Atari models and the Commodore 64. The Apple version preferred additional hardware that contained DACs, although it could instead use the computer's one-bit audio output (with the addition of much distortion) if the card was not present. The Atari made use of the embedded POKEY audio chip. Speech playback on the Atari normally disabled interrupt requests and shut down the ANTIC chip during vocal output. The audible output is extremely distorted speech when the screen is on. The Commodore 64 made use of the 64's embedded SID audio chip (Hetzfeld, 1984).

Atari

Arguably, the first speech system integrated into an operating system was the 1400XL/1450XL personal computers designed by Atari, Inc. using the Votrax SC01 chip in 1983. The 1400XL/1450XL computers used a Finite State Machine to enable World English Spelling text-to-speech synthesis. Unfortunately, the 1400XL/1450XL personal computers never shipped in quantity. The Atari ST computers were sold with "stspeech.tos" on floppy disk (Anderson J. J., 2015).

Apple

The first speech system integrated into an operating system that shipped in quantity was Apple Computer's MacInTalk. The software was licensed from 3rd party developers Joseph Katz and Mark Barton and was featured during the 1984 introduction of the Macintosh computer. This January demo required 512 kilobytes of RAM memory. As a result, it could not run in the 128 kilobytes of RAM the first Mac actually shipped with (Hetzfeld, 1984). So, the demo was accomplished with a prototype 512k Mac, although those in attendance were not told of this and the synthesis demo created considerable excitement for the Macintosh. In the early 1990s Apple expanded its capabilities offering system wide text-to-speech support. With the introduction of faster PowerPC-based computers they included higher quality voice sampling. Apple also introduced speech recognition into its systems which provided a fluid command set. More recently, Apple has added sample-based voices. Starting as a curiosity, the speech system of Apple Macintosh has evolved into a fully supported program, PlainTalk, for people with vision problems. VoiceOver was for the first time featured in Mac OS X Tiger (10.4). During 10.4 (Tiger) & first releases of 10.5 (Leopard) there was only one standard voice shipping with Mac OS X. Starting with 10.6 (Snow Leopard), the user can choose out of a wide range list of multiple voices. VoiceOver voices feature the taking of realistic-sounding breaths between sentences, as well as improved clarity at high read rates over PlainTalk. Mac OS X also includes say, a command-line based application that converts text to audible speech. The AppleScript Standard Additions includes a say verb that allows a script to use any of the installed voices and to control the pitch, speaking rate and modulation of the spoken text.

The Apple iOS operating system used on the iPhone, iPad and iPod Touch uses VoiceOver speech synthesis for accessibility. Some third-party applications also provide speech synthesis to facilitate navigating, reading web pages or translating text (Apple Inc, 2011).

AmigaOS

The second operating system to feature advanced speech synthesis capabilities was AmigaOS, introduced in 1985. The voice synthesis was licensed by Commodore International from SoftVoice, Inc., who also developed the original MacinTalk text-to-speech system. It featured a complete system of voice emulation for American English, with both male and female voices and "stress" indicator markers, made possible through the Amiga's audio chipset. The synthesis system was divided into a translator library which converted unrestricted English text into a standard set of phonetic codes and a narrator device which implemented a formant model of speech generation (Miner, 1991). AmigaOS also featured a high-level "Speak Handler", which allowed command-line users to redirect text output to speech. Speech synthesis was occasionally used in third-party programs, particularly word processors and educational software. The synthesis software remained largely unchanged from the first AmigaOS release and Commodore eventually removed speech synthesis support from AmigaOS 2.1 onward.

Despite the American English phoneme limitation, an unofficial version with multilingual speech synthesis was developed. This made use of an enhanced version of the translator library which could translate a number of languages, given a set of rules for each language (Devitt).

Microsoft Windows

Modern Windows desktop systems can use SAPI 4 and SAPI 5 components to support speech synthesis and speech recognition. SAPI 4.0 was available as an optional add-on for Windows 95 and Windows 98. Windows 2000 added Narrator, a text–to–speech utility for people

who have visual impairment. Third-party programs such as JAWS for Windows, Window-Eyes, Non-visual Desktop Access, Supernova and System Access can perform various text-to-speech tasks such as reading text aloud from a specified website, email account, text document, the Windows clipboard, the user's keyboard typing, etc. Not all programs can use speech synthesis directly. Some programs can use plug-ins, extensions or add-ons to read text aloud. Third-party programs are available that can read text from the system clipboard. Microsoft Speech Server is a server-based package for voice synthesis and recognition. It is designed for network use with web applications and call centers (Microsoft, 2011).

Texas Instruments TI-99/4A

In the early 1980s, TI was known as a pioneer in speech synthesis, and a highly popular plug-in speech synthesizer module was available for the TI-99/4 and 4A. Speech synthesizers were offered free with the purchase of a number of cartridges and were used by many TI-written video games (notable titles offered with speech during this promotion were Alpiner and Parsec). The synthesizer uses a variant of linear predictive coding and has a small in-built vocabulary. The original intent was to release small cartridges that plugged directly into the synthesizer unit, which would increase the device's built in vocabulary. However, the success of software text-to-speech in the Terminal Emulator II cartridge cancelled that plan.

TEXT TO SPEECH SYSTEMS

Text-to-Speech (TTS) refers to the ability of computers to read text aloud. A TTS Engine converts written text to a phonemic representation, then converts the phonemic representation to waveforms that can be output as sound. TTS engines with different languages, dialects and specialized vocabularies are available through third-party publishers (Microsoft, 2007).

Android

Version 1.6 of Android added support for speech synthesis (TTS) (Trivi, 2010).

Internet

Currently, there are a number of applications, plugins and gadgets that can read messages directly from an e-mail client and web pages from a web browser or Google Toolbar, such as Text to Voice, which is an add-on to Firefox. Some specialized software can narrate RSS-feeds. On one hand, online RSS-narrators simplify information delivery by allowing users to listen to their favorite news sources and to convert them to podcasts. On the other hand, on-line RSS-readers are available on almost any PC connected to the Internet. Users can download generated audio files to portable devices, e.g. with a help of podcast receiver, and listen to them while walking, jogging or commuting to work.

A growing field in Internet based TTS is web-based assistive technology, e.g. 'Browsealoud' from a UK company and Readspeaker. It can deliver TTS functionality to anyone (for reasons of accessibility, convenience, entertainment or information) with access to a web browser. The non-profit project Pediaphon was created in 2006 to provide a similar web-based TTS interface to the Wikipedia. Other work is being done in the context of the W3C through the W3C Audio Incubator Group with the involvement of The BBC and Google Inc. (Bischoff, 2007).

Open Source

Systems that operate on free and open source software systems including Linux are various, and include open-source programs such as the Festival Speech Synthesis System which uses diphone-based synthesis, as well as more modern and better-sounding techniques, eSpeak, which supports a broad range of languages, and gnuspeech which uses articulatory synthesis from the Free Software Foundation (Free Software Foundation, Inc., 2015).

Others

1. Following the commercial failure of the hardware-based Intellivoice, gaming developers sparingly used software synthesis in later games. A famous example is the introductory narration of Nintendo's Super Metroid game for the Super Nintendo Entertainment System. Earlier systems from Atari, such as the Atari 5200 (Baseball) and the Atari 2600 (Quadrun and Open Sesame), also had games utilizing software synthesis.
2. Some e-book readers, such as the Amazon Kindle, Samsung E6, PocketBook eReader Pro, enTourage eDGe, and the Bebook Neo.
3. The BBC Micro incorporated the Texas Instruments TMS5220 speech synthesis chip,
4. Some models of Texas Instruments home computers produced in 1979 and 1981 (Texas Instruments TI-99/4 and TI-99/4A) were capable of text-to-phoneme synthesis or reciting complete words and phrases (text-to-dictionary), using a very popular Speech Synthesizer peripheral. TI used a proprietary codec to embed complete spoken phrases into applications, primarily video games.
5. IBM's OS/2 Warp 4 included VoiceType, a precursor to IBM ViaVoice.
6. GPS Navigation units produced by Garmin, Magellan, TomTom and others use speech synthesis for automobile navigation.
7. Yamaha produced a music synthesizer in 1999, the Yamaha FS1R which included a Formant synthesis capability. Sequences of up to 512 individual vowel and consonant formants could be stored and replayed, allowing short vocal phrases to be synthesized.

Digital Sound-alikes

With the 2016 introduction of Adobe Voco audio editing and generating software prototype slated to be part of the Adobe Creative Suite and the similarly enabled DeepMind WaveNet, a deep neural network based audio synthesis software from Google speech synthesis is verging on being completely indistinguishable from a real human's voice (Deepmind, 2016).

Adobe Voco takes approximately 20 minutes of the desired target's speech and after that it can generate sound-alike voice with even phonemes that were not present in the training material. The software obviously poses ethical concerns as it allows to steal other people's voices and manipulate them to say anything desired (Channel, 2016).

This increases the stress on the disinformation situation coupled with the facts that, —

1. Human image synthesis since the early 2000s has improved beyond the point of human's inability to tell a real human imaged with a real camera from a simulation of a human imaged with a simulation of a camera.
2. 2D video forgery techniques were presented in 2016 that allow near real-time counterfeiting of facial expressions in existing 2D video (Thies, 2016).

Speech Synthesis Markup Languages

A number of markup languages have been established for the rendition of text as speech in an XML-compliant format. The most recent is Speech Synthesis Markup Language (SSML), which became a W3C recommendation in 2004. Older speech synthesis markup languages include Java Speech Markup Language (JSML) and SABLE. Although each of these was proposed as a standard, none of them have been widely adopted.

Speech synthesis markup languages are distinguished from dialogue markup languages. VoiceXML, for example, includes tags related to speech recognition, dialogue management and touchtone dialing, in addition to text-to-speech markup.

APIs

Multiple companies offer TTS APIs to their customers to accelerate development of new applications utilizing TTS technology. Companies offering TTS APIs include AT&T, CereProc, DIOTEK, IVONA, Neospeech, Readspeaker, SYNVO, YAKiToMe! and CPqD. For mobile app development, Android operating system has been offering text to speech API for a long time. Most recently, with iOS7, Apple started offering an API for text to speech.

Applications

Speech synthesis has long been a vital assistive technology tool and its application in this area is significant and widespread. It allows environmental barriers to be removed for people with a wide range of disabilities. The longest application has been in the use of screen readers for people with visual impairment, but text-to-speech systems are now commonly used by people with dyslexia and other reading difficulties as well as by pre-literate children. They are also frequently employed to aid those with severe speech impairment usually through a dedicated voice output communication aid.

Speech synthesis techniques are also used in entertainment productions such as games and animations. In 2007, Animo Limited announced the development of a software application package based on its speech synthesis software FineSpeech, explicitly geared towards customers in the entertainment industries, able to generate narration and lines of dialogue according to user specifications. The application reached maturity in 2008, when NEC Biglobe announced a web

service that allows users to create phrases from the voices of Code Geass: Lelouch of the Rebellion R2 characters (Anime News Network, 2007).

In recent years, Text to Speech for disability and handicapped communication aids have become widely deployed in Mass Transit. Text to Speech is also finding new applications outside the disability market. For example, speech synthesis, combined with speech recognition, allows for interaction with mobile devices via natural language processing interfaces.

Text-to speech is also used in second language acquisition. Voki, for instance, is an educational tool created by Oddcast that allows users to create their own talking avatar, using different accents. They can be emailed, embedded on websites or shared on social media (Anime News Network, 2007).

In addition, speech synthesis is a valuable computational aid for the analysis and assessment of speech disorders. A voice quality synthesizer, developed by Jorge C. Lucero et al. at University of Brasilia, simulates the physics of phonation and includes models of vocal frequency jitter and tremor, airflow noise and laryngeal asymmetries. The synthesizer has been used to mimic the timbre of dysphonic speakers with controlled levels of roughness, breathiness and strain (Englert, Madazio, Gielow, Lucero, & Behlau, Perceptual error identification of human and synthesized voices, 2016).

WORKS CITED

Adriaans, F., & Kager, R. (2010). *Adding Generalization to Statistical Learnings: The Induction of Phooactics from Continuous Speech.* Journal of Memory and Language.

Albat, T. F. (2012). *Systems and Methods for Automatically Estimating a Translation Time.*

Allen, J., Hunnicutt, M. S., & Klatt, D. (1987). *From Text to Speech: The MITalk system.* Cambridge: Cambridge University Press.

Anderson, J. J. (2015, February 6). Atari. *Creative Computing*, p. 51.

Anderson, S. R. (2016). Morphology. In S. R. Anderson, *Encyclopedia of Cognitive Science.* New Haven, Connecticut, United States of America: Macmillan Reference Ltd.

Anime News Network. (2007). Speecj Synthesis Software for Anime Announced. *Anime News Network.*

Apple Inc. (2011). *iPhone: Configuring accessibility features (Including VoiceOver and Zoom).* Cupertino: Apple Inc.

Babrow, D. (n.d.). *Natural Language input for a Computer Problem Solving System.* American Association for Artificial Intelligence.

Badecker, W., & Allen, M. (2002). *Morphological Parsing and the Perception of Lexical Identity: A Masked Priming Study of Stem Homographs.* Journal of Memory and Language.

Bischoff, A. (2007). PDA's and MP3-Players, Proceedings of the 18th International Conference on Database and Expert Systems Applications. In A. Bishchoff, *The Pediaphon - Speech Interface to the free Wikipedia Encyclopedia* (pp. 575-579).

Black, A. W. (2002). *Perfect synthesis for all the people all of the time.* IEEE TTS Workshop.

Camargo, J. E., & Gonzalez, F. A. (2009). A Multi-class Kernel Alignment Method for Image Collection Summarization. In Proceedings of the 14th Iberoamerican Conference on

Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. In J. E. Camargo, & F. A. Gonzalez, *Eduardo Bayro-Corrochano and Jan-Olof Eklundh.* Berlin: Springer-Verlag.

Channel, B. B. (2016, November 7). *Adobe Voco 'Photoshop-for-voice causes concern*. Retrieved from British Broadcasting Channel: BBC.com

Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., & Tsujii, J.-i. (2006). *Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning.* Pacific Sympsium on Biocomputing.

Church, K. W. (1988). A stochastic parts program and noun phrase for unrestricted text. *Proceedings of the second conference on Applied natural language processing. Association for Computational Linguistics.* Stroudsburg: ANLC.

Dagan, I., Clickman, O., & Magnini, B. (2004). *Probabilitic Textual Entailment: Generic applied modelling of language variability.* Grenoble: PASCAL Workshop on Learning Methods for Text Understanding and Mining.

Dartmouth College. (1993). *Musical and Computers.* Wayback Machine.

Dartmouth College. (2011, June 08). Music and Computers. *Wayback Machine*. Hanover, New Hampshire, United States.

Deepmind. (2016, September 08). *A Generative Model of Raw Audio*. Retrieved from WaveNet: Deepmind.com

Devitt, F. (n.d.). *Translator Library (Multilingual-speech version).*

Dictionary, M. (2012). *British English definition of voice recognition.* Macmillan Publishers Limited.

Discourse Analysis - What Speakers Do in Conversion . (2016, February 20). Retrieved from Linguistic Society of America: www.linguisticsociety.org

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Vreckren, O. v. (1996). *The MBROLA Project: Towards a set of high quality speech synthesizers of use for non commercial purposes.* ICSLP Proceedings.

Elaine, M., & Perzanowski, D. (1998). *MUC-7 Evaluation of IE Technology: Overview of Results.*

Englert, M., Madazio, G., Gielow, I., Lucero, J., & Behlau, M. (2015). *Perceptual error identification of human and synthesized voices.* Journal of Voice.

Englert, M., Madazio, G., Gielow, I., Lucero, J., & Behlau, M. (2016). Perceptual error identification of human and synthesized voices. *Journal of Voice*.

FGC/SRS Speech Recognition Server. (2007). *Fifth Genration Computer Corporation*. Retrieved from FifthGen.com: http://www.fifthgen.com

Freddy, Y. Y. (2000). Advances in domain independent linear text segmentation. *1st Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 26-33). ANLP-NAACL.

Free Software Foundation, Inc. (2015, February 18). *GNU Operating System*. Retrieved from GNU Operating System: http://www.gnu.org

Green, N., Breimyer, P., Kumar, V., & Samatova, N. F. (n.d.). *WebBANC: Building Semantically-Rich Annotated Corpora from Web.* Raleign, North Carolina, United States of America: North Carolina State University.

Hetzfeld, A. (1984, January). *It Sure Is Great To Get Out Of That Bag!* Retrieved from Folkore: www.Folklore.org

Juscyk, P. W., & Houston, D. M. (1999). *The Beginnings of Word Segmentation in English-Learning Infants.* Cognitive Psychology.

Kishorjit, N., Rai, R. K., & Sivaji, B. (2012). Manipuri Morpheme Identification. *3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)* (pp. 95 - 108). Mumbai: COLING.

Lamel, L. F., Gauvain, J. K., Prouts, B., Bouchier, C., & Boesch, R. (1993). *Generation and Synthesis of Broadcast Messages.* Proceedings ESCA-NATO Workshop and Applications of Speech Technology.

Lieberman, H., Faaborg, A., Daher, W., & Espinosa, J. (2005). *How to wreck a Nice Beach You Sing Calm Incense.* Cambrige, MT: MIT Media Library.

Lovins, J. B. (1968). *Development of a Stemming Algorithm". Mechanical Translation and Computational Linguistics.*

Microsoft. (2007). *How to configure and use Text-to-Speech in Windows XP and in Windows Vista.* Microsoft.

Microsoft. (2011). *Accessibility Tutorials for Windows XP: Using Narrator.* Microsoft.

Milligan, M. (2016). UPHE Snags Exclusive 'Barbie' SVOD Rights from Mattel. *Animation Magazine*.

Miner, J. (1991). *Amiga Hardware Reference Manual.* Addison-Wesley Publishing Company, Inc.

Morante, R., Krallinger, M., Valencia, A., & Daelemans, W. (CLEF). *Machine eading of Biomedical Texts about Alzheimer's Disease.* 2012: Evaluation Labs and Workshop.

Moro, A., Raganato, A., & Navigli, R. (2014). *Entity Linking meets Word Sense Disamguation: a Uified Approach.* Transactions of the Association for Computational Linguistics.

Oura, K. (2017, January 16). *HMM-based Speech Synthesis System*. Retrieved from HTS: http://hts.sp.nitech.ac.jp/

Park, Y., Byrd, R. J., & Boguraev, B. (2002). Automatic glossary extraction: beyond terminology identification. *Computational Linguistics, 19th International Conference on Computational lingistics.* Taipei.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambrige: MIT Press.

R, P., & P, N. (n.d.). Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature. In P. R, & N. P, *Computing and Informatics* (pp. 1-32).

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). *Speech perception without traditional speech cues.* Science.

Rubin, P., Baer, T., & Mermelstein, P. (1981). *An articulatory synthesizer for perceptual research.* Journal of the Acoustical Soceity of America.

Santen, J. P., Sproat, R. W., Olive, J. P., & Hirschberg, J. (1997). *Progress in Speech Synthesis.* New York: Springer-Verlag.

Schantz, H. F. (2012). *The history of OCR, Optical Character Recognition.* Manchester: Recognition Technologies Users Association.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2009). *Automatic Extraction of Rules for Sentence Boundary Disambiguation.* University of Patras.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A Compuer Approach to content analysis.* Cambridge: MIT Press.

Taylor, P. (2009). *Text-to-speech synthesis.* Cambridge, UK: Cambridge University Press.

Thies, J. (2016, June 18). Face2Face: Real-time Face Capture and Reenactment of RGB videos. *Computer Vision and Pattern Recognition*.

Trivi, J.-M. (2010, February 17). *An introduction to Text-to-Speech in Android*. Retrieved from Android Developers: Android-developers.blogspot.com