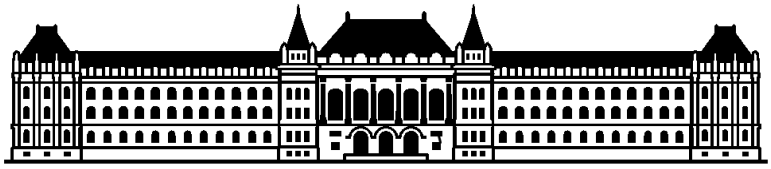

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
DOCTORAL SCHOOL OF INFORMATICS
DEPARTMENT OF TELECOMMUNICATIONS AND MEDIA INFORMATICS

THE FUNCTIONAL EXPANSION OF AUTOMATIC SPEECH RECOGNITION OUTPUT BASED ON NOVEL PROSODY- AND TEXT-BASED APPROACHES

Ph.D. thesis

Máté Ákos Tündik

M.Sc. in Computer Science Engineering

Supervisor:

György Szaszák, Ph.D.

BUDAPEST, HUNGARY
2019

Alulírott, Tündik Máté Ákos kijelentem, hogy ezt a doktori értekezést magam készítettem és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

A dolgozat bírálatai és a védésről készült jegyzőkönyv a későbbiekben, a Budapesti Műszaki és Gazdaságtudományi Egyetem dékáni hivatalában lesz elérhető.

Budapest, 2019. október 20.

.....

Kivonat

Napjainkban a beszédfelismerő rendszerek ipari alkalmazása egyre elterjedtebbé vált; jelenleg a beszédtechnológia elsődleges céljai közé tartozik az emberi nyelven történő kommunikáció automatikus/gépi értelmezése, az ember számára lényeges információ kinyerése. Azonban mivel a beszédfelismerők puszta szó- vagy karaktersorozatot adnak vissza, annak értelmezése, strukturális tagolása nélkül, fontos, hogy ezt a szöveges kimenetet minél inkább funkcionálisan bővítsük. Ehhez mind a beszédjel (akusztikum) közvetlen feldolgozását megcélzó technikák, mind a szöveg feldolgozását támogató NLP (Natural Language Processing) módszerek jelentősen hozzájárulnak.

A disszertáció egyik része a beszédprozódia területén belül a hangsúlydetektálás vizsgálatára irányul, melynek segítségével fonológiai frázisdetektálás készíthető. A fonológiai frázisokra történő szegmentálás a beszédjelből származó információstruktúrát hivatott tükrözni. A célom az volt, hogy különböző nyelvek (francia és magyar) kötött hangsúlyozási szabályszerűségeihez adaptálható modellt dolgozzak ki, egy meglévő intonáció-modellező algoritmusra építve. Algoritmusomat egy HMM/GMM alapú megoldással kombinálva hibrid rendszert is implementáltam, mely a különálló módszerek teljesítményét meghaladta.

Doktori témámban nemcsak a beszédjel strukturálhatóságát vizsgáltam meg, hanem a beszédfelismerésből származó központozatlan szöveges kimenet írásjelekkel történő bővítési lehetőségét is. Ehhez rekurrens neurális hálózatokon alapuló modelleket hoztam létre. A szövegek írásjelekkel történő strukturálása nemcsak a gépi beszédértelmezésben játszik fontos szerepet. A dolgozatomban szubjektív vizsgálatokon keresztül mutatom be, hogy nemcsak a kézzel, hanem (bizonyos hibakorlát mellett) az automatikusan központozott feliratok is olvashatóbbak, érthetőbbek az emberek számára, a központozatlan változatokhoz képest. Téziseimet egy nagy létszámú kontroll-csoport, valamint egy kisebb, siket és nagyothalló emberekből álló csoport bevonásával igazoltam.

Kulcsszavak

Az alábbi kulcsszavak indexelési célokat szolgálhatnak:
automatikus beszédfelismerés, ASR, prozódiai hangsúlydetektálás, fonológiai frázisdetektálás, atom dekompozíció, hibrid modell, zárt feliratozás, információgazdag átirat, írásjel-visszaállítás, rekurrens neurális hálózatok, konvolúciós neurális hálózatok, szubjektív tesztek, automatikus kivonatolás, információkinyerés, hibaterjedés

Abstract

Nowadays, the industrial application of speech recognition systems has become increasingly widespread. One of the primary purposes of speech technology is to automatically understand the meaning of the spoken language, by extracting the relevant information from the human speech. However, since automatic speech recognizers (ASRs) provide often just a raw output (basically, a continuous word- or character-sequence) without any further interpretation, in this case, it is important to expand and structure this text as functionally as possible. To achieve this, two different kinds of technique can contribute significantly; the direct processing of the speech signal (acoustics) and NLP (Natural Language Processing) with text input.

The first part of my thesis is aimed at a speech prosody-related automatic segmentation method, reflecting the information structure; (i) First, detecting the stress in the speech signal, (ii) which can be used to make phonological phrase (PP) detection. My goal was to propose models adapted to the stress characteristics of different fixed stressed languages (French and Hungarian), based on an existing intonation modelling algorithm. Combining the novel PP method with a HMM/GMM baseline, I implemented a hybrid solution that exceeded the performance of the base approaches.

In my thesis, I not only studied the structure of the speech signal, but also the possibility of expanding the raw ASR output with automatically inserted punctuation marks. To achieve this, I created recurrent neural network-based models. Structuring the texts with punctuation plays an important role not only in automatic spoken language understanding. In my thesis, I show through subjective examinations that not only the manually, but (with a certain margin of error) the automatically punctuated texts are more readable and understandable for humans than the pure word chain without punctuation. My theses were verified by involving a large group with hearing people, and a small group of deaf and hard-of-hearing (DHH) people.

Keywords

The following keywords may be useful for indexing purposes:
automatic speech recognition, ASR, prosodic stress detection, phonological phrasing, atom decomposition, hybrid model, closed captioning, rich transcription, punctuation restoration, sequence-to-sequence model, recurrent neural networks, convolutional neural networks, subjective tests, automatic summarization, information extraction, error propagation

I dedicate this dissertation
to the Almighty God and my Beloved Ones,
keeping alive myself during every second of my research period.

Köszönetnyilvánítás

2015-ben döntöttem úgy, hogy elkezdem a Ph.D.-képzést; mindvégig tudtam, hogy ezt munka mellett nem lesz egyszerű véghezvinni. De aki ismer, az tudja, hogy teljes erőbedobással, nyughatatlanul küzdök, amíg el nem érem a célt. A Ph.D.-hoz vezető fárasztó, de gyümölcsöket is termő út mellett legalább annyira fontos az az önismereti utazás, amit megtettem; megélni a mélységet, majd újra és újra erőre kapni, és örülni, ha értelmet nyert a munkám.

Ezen az úton sokan elkísértek, mint szakmai és/vagy lelki támogatók. Elsőként külön köszönöm Dr. Szaszák Györgynek (Gyurinak), hogy a B.Sc. és az M.Sc. után harmadszor is vállalta témavezetőként azt a "kockázatos projektet", hogy velem kellett együtt dolgozni. Mindamellett, hogy szakmai tanácsokkal segített a publikációimban, egyetemi szervert biztosított a kísérleteimhez, a kutatásaimhoz tartozó pénzügyi támogatásért is köszönettel tartozom neki. Köszönöm a BME TMIT Beszédakusztikai Laboratórium munkatársainak, Dr. Vicsi Klárának, Nagy Ildikónak, Lénártné Kovács Annamáriának, Dr. Sztahó Dávidnak, Kiss Gábornak, és Tulics Miklós Gábrielnek az elmúlt évek közös munkáit. Köszönet illeti a SpeechTex Kft. dolgozóit is; Fegyó Tibort, Mihajlik Pétert, és nem utolsó sorban Tarján Balázst, akivel élmény volt a szakmai eszmecsere.

Köszönöm a Beszédkommunikáció és Intelligens Interakciók Laboratórium munkatársainak, Dr. Németh Gézának, Dr. Olaszy Gábornak, Dr. Zainkó Csabának és Dr. Csapó Tamás Gábornak a téziseimmel kapcsolatos visszajelzéseket. Köszönöm a BME TMIT összes dolgozójának, külön kiemelve Dr. Magyar Gábor Tanszékvezető Úr, Dr. Bíró József, Dr. Halász Edit és Dr. Sallai Gyula támogatását. Dr. Németh Krisztiánnak külön köszönöm a biztatását és a Ph.D.-teendőkben nyújtott segítségét (és persze a kávékat).

Köszönöm barátaimnak, hogy az együtt és külön töltött évek alatt elviseltetek. Lóri, H. Attila, M. Attila, Bálint és Ági, Peti, Andris és Kriszti, G. Attila és Ági, Isti és Noémi, és mindenki: hálás vagyok Nektek! Külön köszönöm Attilának, lelkipásztoromnak és jóbarátomnak a rengeteg imádságát!

Köszönöm a Családomnak, hogy féltettek...de ahogy mindig, elfogadták a tervemet, és velem voltak: Anya, Apa, Ágoston és Marci. Mama, sajnos nem lettem orvos, de remélem annak is fogsz örülni, hogy ha Ph.D. leszek.

Végezetül: hálás vagyok Páromnak, Briginek: Mindenért. Amit más nem látott és nem hallott, te mindig érezted és tudtad. Fogadd millió köszönetemet.

Acknowledgement

In 2015, I decided to start the Ph.D.-school; I knew all along that it wouldn't be easy to do it besides having a job. But anyone, who knows me, also knows that I struggle with full force until I reach my goal. In addition to the tiring but fruitful journey to the Ph.D., the journey of self-knowledge I have made is also important; to live the depths, then to build yourself up over and over again, and finally to feel happy about my work.

So many people have accompanied me along this path as professional and / or emotional supporters. First of all, I would like to thank Dr. György Szaszák (Gyuri) that for the third time after the B.Sc. and M.Sc. he took on the "risky project" of working with me. In addition to providing me technical advices on my publications, ensuring me a university server for my experiments, I also thank him for his financial support for my research. I would like to thank the colleagues of BME TMIT Speech Acoustics Laboratory, Dr. Klára Vicsi, Ildikó Nagy, Annamária Lénártné Kovács, Dr. Dávid Sztahó, Gábor Kiss, and Gábor Gabriel Tulics for their joint work.

I also thank to the employees of SpeechTex Kft .; Tibor Fegyó, Péter Mihajlik, and last but not least Balázs Tarján, with whom I loved the professional conversations. I would like to thank the staff of the Speech Communication and Intelligent Interaction Laboratory, Dr. Géza Németh, Dr. Gábor Olaszy, Dr. Csaba Zainkó and Dr. Tamás Gábor Csapó, for their feedback on my thesis.

I would like to thank all the employees of BME TMIT, especially the support of Dr. Gábor Magyar Head of Department, Dr. József Bíró, Dr. Edit Halász and Dr. Gyula Sallai. Special thanks to Dr. Krisztián Németh for his encouragement and help with my Ph.D.-tasks (and of course, the coffee).

Thank you my friends for having endured me through years, together and separately. Lóri, Attila H., Attila M., Bálint and Ági, Peti, Andris and Kriszti, Attila G. and Ági, Isti and Noémi, and everyone: I am grateful to you! Special thanks to Attila, my pastor and my good friend for the many prayers!

Thank you to my family to worried about me ... but as always, they accepted my plan and were with me: Mom, Dad, Ágoston and Marci. Mom, unfortunately I didn't become a doctor, but I hope if you see me becoming a Ph.D., you will be happy.

Finally, I am grateful to my partner, Brigi: for everything. What others did not see or hear, you always felt and knew. Receive a million thanks.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

The quote "The only constant thing is change", said by Heraclitus, is totally valid for the area of speech technology as well. Regarding automatic speech recognition (ASR) which became the part of a wide range of applications in our everyday life, historically it was restricted to speech-to-text transformation, without any further analysis such as meaning extraction. In the Big Data and artificial intelligence era, this is not sufficient any more. The researchers got interested in whether a machine can behave similar to humans, or have similar level of intelligence like a human being [1]. Naturally, this way of thinking has affected speech technology as well. The ASR output is further processed, often referred to as Spoken Language Understanding (SLU), where a machine is able to detect the meaning and the intent from speech utterances. SLU incorporates various tasks like domain classification, user intent detection, slot filling, often demonstrated through a personal assistant example [2, 3]. SLU requires intelligent automatic processing of speech and text. For text, the usage of Natural Language Processing (NLP) modules that perform syntax and semantics related tasks like part-of-speech (POS) tagging, Named Entity Recognition (NER) or dependency parsing could be straightforward, but they are originally designed for error-free transcripts. Although the ASR output is rarely 100% correct, making the used NLP approaches word error robust has received very little attention. On the other hand, acoustic cues besides the textual ones can be exploited for these purposes of SLU. In my dissertation, I focus on these cues, in order to provide important functional expansion for the ASR output, with automatic structural segmentation methods; punctuation and phrasing.

It is known that suprasegmental (prosodic) features can help ASR-systems by providing

information structure-related cues (sentence, syntactic phrase or word boundaries) and also cues related to emotions [4]. The detection of the acoustic markers can improve not only the robustness of the ASR-systems [5], but they could be well-exploited for SLU-tasks as well, i.e. by enriching the raw word sequence (output transcription) with some annotation, such as punctuation marks or capitalization [6].

Besides the changing trends from speech recognition to speech understanding, the rapidly changing methodology is also worth to mention. The appearance of Deep Learning, especially Deep Neural Networks (DNNs) has reformed the entire area of speech technology. According to the latest Gartner's hype curve (2018), DNN is on the "Peak of Inflated Expectations"; many business and research projects aim at exploiting its advantages, often using Big Data analytics. A nice property of the DNN-based models is that they allow for capturing and modelling multiple level of abstractions of speech and text through numerous layers of different architectures [7], but still based on the same paradigm (and general framework). For instance, acoustic modelling has changed from HMM-GMM hybrids to HMM-DNN hybrids [8], then nowadays all neural (end-to-end) approaches are considered [9]. In text processing, a revolution came from the Word2Vec method [10], which can be also used for language modelling. DNN-based end-to-end solutions can be found for SLU in the literature as well [11].

Hungarian language – due to its heavily agglutinating nature and relatively free word order, and relative small number of native speakers compared to world languages – has always been a special case of ASR and NLP research within Hungary. The application of HMM-DNN hybrid systems for ASR is also well-studied [12, 13], and in NLP, different approaches for POS-tagging and NER were also evaluated [14, 15, 16, 17, 18, 19], or even connected into a syntactic-semantic processor pipeline [20]. In SLU, combining text-based techniques and acoustic analysis, keyword spotting [21], semantic focus detection [22] and automatic summarization [23] methods have been investigated.

1.2 Outline of the Thesis

The thesis is organized as follows.

Chapter 2 describes the relevance of stress detection from speech prosody, and the related techniques. Based on stress detection, phonological phrasing is applicable. First I demonstrate a baseline HMM/GMM-approach for this purpose, then my new WCAD-based stress detection and phrasing solution is presented, followed by a hybrid approach, discussing the new theses proven by experiments for Hungarian and French. I also mention an example where the phonological phrasing methods can be applied in practice.

Chapter 3 reflects on the importance of punctuation, and discusses the different automatic punctuation paradigms. Then a baseline Maximum Entropy-based model for this purpose is introduced. After that, I present novel word- and character-based approaches for Hungarian, applying recurrent neural networks. Finally, with the help of these concepts and an acoustic-prosodic extension, I establish hybrid solutions. I also make a comparative assessment between Hungarian and English results for the automatic punctuation task.

Chapter 4 presents a subjective, user-centric evaluation of ASR-transcripts, with particular attention to automatically punctuated texts. To the best of my knowledge this is a unique effort as I was not able to find similar subjective assessment for automatic punctuation from end-user perspective[1]. A large group of hearing people, and a small group of deaf and hard-of-hearing (DHH) people were involved in the experiments, then statistical examinations were performed, concluding some important theses.

Chapter 5 provides a short overview of my theses, emphasizing the most important conclusions, and some future directions for further improvements.

Chapter 6 demonstrates the applicability of my results.

[1]This unique effort was confirmed by the Best Student Paper Jury of Interspeech 2018 conference

Chapter 2

Atom decomposition-based Prosodic Stress Detection and Phonological Phrasing

2.1 Background

2.1.1 Terminology

In the following sections, I present speech prosody-based automatic segmentation solutions, which are important for Spoken Language Understanding. Speech prosody plays an important role in human speech perception and also in speech technology applications. Prosody reflects the hierarchical structure of the language [24], and conveys information linked to human discourses [25] and emotions [26]. Without going into any further details about describing prosody, starting from elementary level, I prefer to focus only on the terms and definitions relevant for my theses, such as stress, accent and phonological phrase. The reader can find abundant literature on the theory of prosody.

2.1.1.1 Disambiguation between Stress and Accent

Unfortunately, the terminology for my research area is not uniform in the literature, differences exist within and between linguistic and engineering approaches. One can read many definitions for the terms *stress*, *emphasis*, *prominence*, *salience*, *accent*, etc., and the distinction among them is not clear. The terminology is often confused, and even a conceptual mismatch is often encountered related to this topic. For instance, the terms *pitch accent*

and *pitch tracker* illustrate this mismatch between the linguistic and psycho-acoustic (physical) approach. To be concrete, pitch is a perceptional and hence subjective notion, as only the fundamental frequency (F0) as physical parameter can be measured. Therefore, strictly speaking, we should use the term F0 tracker, which is not the case in the literature.

To eliminate the possible ambiguities, in the following chapters, I use the term stress as covering any perceptually or acoustically relevant prominence or salience in any physically measurable objective parameter (F0, intensity, duration, etc.). Additionally, I use the term accent to denote F0 movements which can be of either positive (high accent) or negative (low accent) direction. As we will see, in the investigated languages (Hungarian and French), a high accent often correlates with stress, whereas low accents are often precursors of stress on the next syllable or follow a stressed syllable.

Stress is often differentiated on phrase- or sentence-level (referred to as *prosodic stress*, in which more words are likely to be emphasized), and on word-level (referred to as *lexical stress*) [27]. Again here, a mismatch occurs whether we are speaking about levels in the phonological or in the syntactic sense.

2.1.1.2 Phonological Phrase

It is known, that syntax is organized layerwise [28], where the N^{th} higher level linguistic component is composed of one or more $N-1^{th}$ level constituent(s). From a bottom-up view, the phonemes construct morphemes, which can form words. The words are aggregated to syntactic phrases, which finally are organized into sentences.

Similarly, there is a prosodic hierarchy described by phonology [29, 30, 31], where the following layers or levels are differentiated (also from bottom-up view); starting from the syllables and feet, the smallest meaningful unit is the phonological word, which is the building block of the phonological phrase. Phonological phrases constitute a prosodic unit characterized by a single, acoustically marked stress and some preceding/following intonation contour by definition. They belong one level below the better known intonational phrase level in the prosodic hierarchy. Finally, the intonational phrases form utterances.

Levelt's speech production model helps to establish a relationship between these hierarchies [32], that is the syntax/phonology interface. In our brain, the sentences are constructed from intonational phrases, which – as I noted – are composed of phonological phrases. According to Selkirk [30], the syntax-prosody mapping is possible with the alignment of syntactic phrases and phonological phrases [33, 34].

It is noted again, that a phonological phrase (PP) by definition contains one and exactly

one stressed position. In practice (i.e. for speech technology applications requiring prosodic annotation), best is to ensure that PP boundaries are always located at word boundaries. In this case not every word boundary is a PP boundary, but a PP boundary is always also a word boundary.

2.1.2 About Stress Detection and Phrasing

Both automatic stress detection and phrasing belong to the important, open problems in SLU. Prosodic cues can be effectively leveraged to address this problem. Automatic phrasing has a dedicated importance before subsequent processing of ASR output, as it establishes a segmentation of the speech stream. The automatically detected intonational units are suitable for various classification tasks such as topic segmentation, labelling dialogue acts, parsing, information extraction, meeting summarization and speech understanding in general [35, 36, 37]. The investigation of acoustically marked stress events linked to SLU have been also started decades ago, such as analysing correlation with word importance or topic introduction [38, 39].

Due to the different language characteristics, neither automatic stress detection, nor phrasing has universal methods. Nevertheless, there is a common point, as the main, measurable suprasegmental features, such as fundamental frequency (F0), energy, and the duration (related to words or phonemes, or pauses), or their changes/movements are universally exploited [40]. Grouping the languages regarding rhythm [41], stress-timed and syllable-timed languages are described. In stress-timed languages such English or Russian, the meaning of the same multi-syllabic words may depend on the stressed syllable (this is lexical stress, i.e. often between verbal and noun forms). For these languages, lexical stress detection on word or on syllable-level are popular [42]. Furthermore, it is a typical use case to integrate these solutions into computer-aided pronunciation learning (CAPL) tools for L2 learning, for example to detect the mispronounced words effectively [43, 44]. Nowadays, the deep learning-based models also became popular in this area; the authors of [43] used a deep feedforward neural network (DFN) and a convolutional neural network (CNN) with a set of temporal and spectral features related to the duration, F0 and energies in different frequency bands over syllable nucleus, for English and Arabic languages. Other languages, like Romanian and Italian belong to the "syllable-timed" group, while Japanese language seems monotone compared to European languages; in these cases, besides the pause duration, the detection of F0 resets [45] and F0 rising [46] can locate the phrase boundaries.

In my thesis, I am studying automatic stress detection and phrasing for Hungarian and

French. French has a fixed stress (mainly on the last syllable [47]), and Hungarian also has a fixed stress, but on the first syllable [48]. Fixed stress makes the related detection task somewhat easier and allows for addressing phrase segmentation (phrasing) and stress detection at the same time, as stressed syllables are adjacent to word boundaries. The authors of [49] reached 93% accuracy in a binary classification task, where a CNN approach assigns stress to words on a French spoken language corpus, based on spectral and temporal filterbank features. Martin designed a complex rule-based automatic phrasing system for French [50]. He showed that the syllable candidates for stress can depend on the speech rate, the part-of-speech (noun, verb), the duration of silence and the F0 change. However, the whole approach was not evaluated with objective metrics. Christodoulides et al. organized a perception test among naive and expert listeners in French [51]; a database from the manually marked prosodic phrase boundaries was built, then an automatic annotation approach was developed. The key predictors for boundary detection were the silent pauses following the boundary syllable, the relative mean F0 and relative duration of two previous syllables, similarly to [50]. The automatic algorithm detected the main prosodic boundaries by 92% F1-score. For Hungarian, Czap and Pintér studied the detection of prosodic stress relying on the energy differences measured on vowels, within syllables [27]. An HMM/GMM approach was presented in [5], which uses a Viterbi alignment to obtain phonological phrase boundaries, relying on F0 and energy features (see details in 2.3).

I explore the possibility of prosodic stress detection and automatic phrasing, then compare my method to the aforementioned baseline HMM/GMM approach [5]. To achieve this, I use an intonational model called Weighted Correlation-based Atom Decomposition (WCAD) algorithm [52] to detect stress (and PP boundaries). Intonation models have been studied exhaustively, especially in the context of speech synthesis ([53] for Hungarian). Recently, the physiologically inspired Generalised Command Response (GCR) model has been proposed [54] which models the F0 contour by decomposing it into elementary atoms using WCAD [52]. In the following chapters, I prove the usability of the WCAD algorithm to automatic prosodic stress (onward referred to simply as stress) and phrase boundary detection based on the obtained atom decomposition in Hungarian and French. I also propose a hybrid model for phonological phrasing which consists of my physiologically inspired WCAD-based approach and the phonologically inspired HMM/GMM component from [5]. As I mentioned, this latter phonological phrase segmentation system is used as a baseline to evaluate my WCAD-based approach for Hungarian and French.

In the international literature (especially in English), ToBI-labelled [55] corpora often

serve for benchmarking these tasks [56]. I skipped this solution, because ToBI does not have a Hungarian adaptation, indeed, attempts to adapt ToBI for Hungarian concluded that no consensus was reached [57] for basic guidelines what and how to annotate in Hungarian prosody. In [58] the difficulties behind this lack of consensus are also illustrated with examples. Instead of producing the ToBi-annotation for the involved corpora, I addressed a qualitative evaluation of the relations between the detected atoms and the syllable structure, i.e. how prosodic stress is linked to syllables in the involved two languages, Hungarian and French. Based on these characteristics, a new phonological phrasing algorithm can be conceptualized.

2.2 Speech Materials

I use read speech utterances which contain individual sentences or short passages composed of 5–10 sentences for my experiments. I prefer to preserve the coherent sentence context wherever possible to obtain authentic prosody. Most of the utterances are taken from newspapers or contemporary literature such that a phonetically balanced corpus is created. The Hungarian corpus contains utterances from 60 speakers (2.8 hours), selected from Hungarian BABEL corpus [59]. The French SIWIS corpus was recorded from 20 speakers (2.2 hours) [60]. However, only a subset of these corpora is segmented at PP-level; hand-labelled references are provided for 400 sentences in case of BABEL, while for 300 sentences in case of SIWIS. Word and phone annotations are generated via a forced alignment with an Automatic Speech Recognizer (ASR), then a rule-based syllabifier is executed for the Hungarian and French corpus as well.

As the segmentation for PPs (illustrated in Fig. 2.1 for French) is based on perception and is to some extent subject to variability, different annotators may annotate the same utterance differently, and both segmentations can be regarded as correct. Following the consensus on the PP set to be used, inter-annotator agreement for Hungarian was 86.2% on a random 10% subset of the annotated data. For French, inter-annotator agreement on PP boundaries was 84.1%, measured on the subset used to obtain consensus on the PP set, corresponding to 6% of the labelled data.

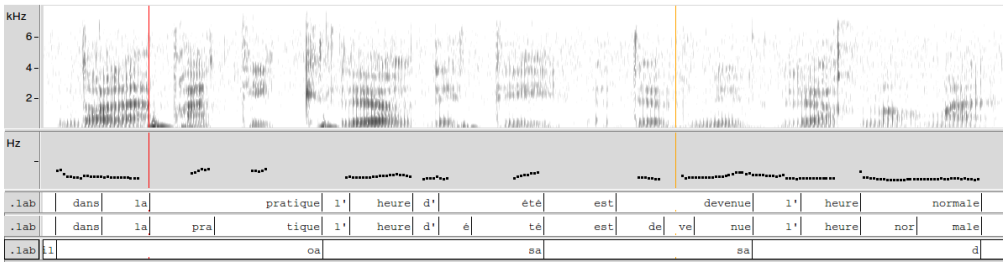

Figure 2.1: Example word, syllable and PP annotation for the utterance snippet 'Dans la pratique l'heure d'été est devenue l'heure normale. [In practice, summertime has become normal time.]"

2.3 The Baseline HMM/GMM-based Phonological Phrasing Approach

The HMM/GMM based system exploits a phonological phrase (PP) model inventory and performs automatic alignment of PPs. In this section I briefly present this system described in detail in [5] and [61]. This system constitutes my baseline for the experiments to be presented. The motivation for selecting this baseline has been discussed in Section 2.1.2.

11-state left-to-right HMM/GMM models have been used, which require fundamental frequency and wide-band energy as acoustic-prosodic features, as well as their first and second order derivatives. Duration features are not exploited directly as within the applied HMM framework, adding them to the feature set did not result in improvement either for Hungarian [61] or for French. For machine learning of the PP models, 300-300 sentences with hand-labelled phonological phrases were provided from the previously presented databases in both languages, resulting in approx. 2k Hungarian and 1.5k French PP examples.

Prosodic stress detection is obtained after the automatic PP alignment, exploiting also the fixed stress: as it was mentioned, in Hungarian, each first syllable of a phrase, in French, each last syllable of a phrase are candidates for stress.The PP-segmentation of the utterances is obtained via a Viterbi-alignment. For fixed stress languages, if stress occurs on the edge, phonological phrasing and stress detection can be regarded as almost equivalent tasks. This means that in Hungarian, where stress is fixed on the first syllable (left edge), each PP onset is also the onset of a stressed syllable. For French, the stress can be regarded as fixed on the last syllable (right edge), although this infers with contours expressing modality. Whereas in Hungarian a first syllable stress is less influenced by intonational constraints, in French, a more complex behaviour of these patterns can be observed at the right edge of the phrases [62]. Differences between PPs are assessed based on the strength of the stress

and the intonation contour of the PP. The PP types used are presented in Table 2.1 for Hungarian and in Table 2.2 for French.

Table 2.1: Modelled PP types for Hungarian

Label	Stress	Intonation contour
od	strong	Clause onset then descending
sd	strong	Stress then descending
ms	medium	Stress then descending
de	medium	Stress then low ending
cr	medium	Stress then ascending
ls	neutral	No stress, descending
sil	neutral	silence

Table 2.2: Modelled PP types for French

Label	Stress	Intonation contour
oa	strong	Clause onset then ascending
sa	strong	ascending, then accent
ma	medium	slight descent, then accent
ne	neutral	balanced
de	neutral	descending
sil	neutral	silence

During the Viterbi-alignment, all PPs are allowed to occur with equal probability by a simple unigram PP grammar [61]. A parameter $logP_{ins}$ influencing insertion likelihood for PPs can be tuned to force or prevent a denser segmentation for PPs. The denser alignment is required, the higher the probability of inserting false PP boundaries becomes, resulting often from a confusion between micro-prosodic variations and accents induced by stress. This approach is documented in details in [61], and is illustrated later in Fig. 2.8.

2.4 Weighted Correlation based Atom Decomposition (WCAD) for F0 Contour Reconstruction

The Weighted Correlation based Atom Decomposition (WCAD) method [52] is an intonation model that describes the F0 contour as the superposition of a global phrase atom and local accent atoms given by the gamma probability distribution function:

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-t/\theta} \quad \text{for} \quad t \geq 0\,. \tag{2.1}$$

In Equation 2.1, θ is the scale parameter which defines the width of the atom, k is the shape parameter which corresponds to the system order, and $\Gamma()$ is the gamma function. The gamma probability distribution function and the gamma function are shown in Figure 2.2.

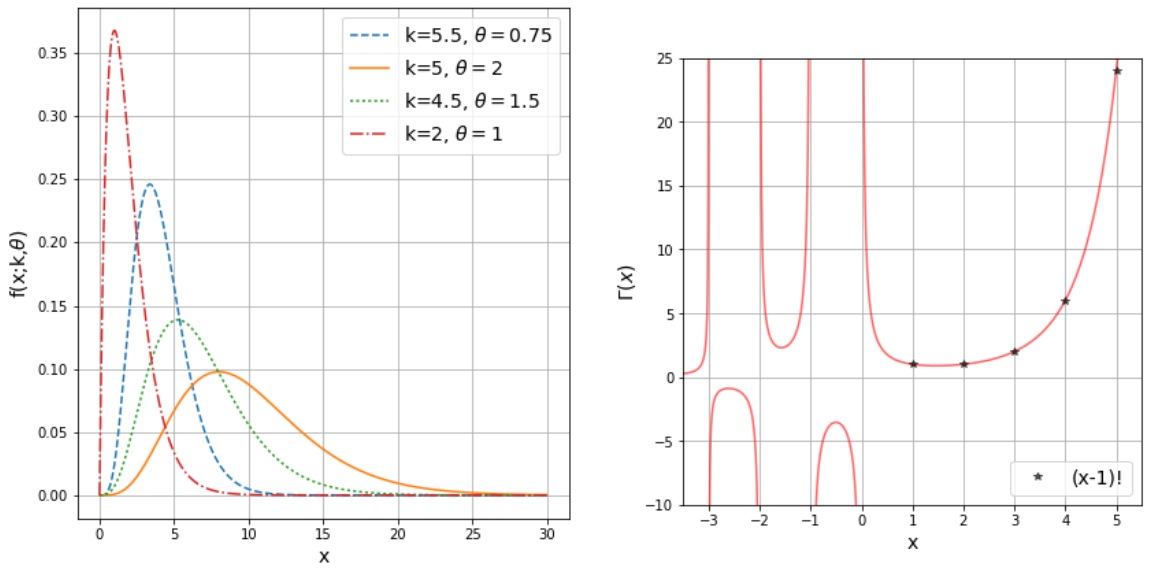

Figure 2.2: The gamma probability distribution function (left) and the gamma function (right)

The gamma distribution function is used for its shape which is regarded to be the approximation of F0 (or eventually smoothed mean energy) movements on stressed syllables.

The elementary atoms are extracted from the F0 contour in the log frequency domain using the Orthogonal Matching Pursuit (OMP) framework [63], [64].The OMP was modified to use the perceptually significant weighted correlation (WCORR) [65] as a cost function, given by:

$$r = \frac{\int w(t)\hat{f}_0(t)f_0(t)\,\mathrm{d}t}{\sqrt{\int w(t)\hat{f}_0(t)\,\mathrm{d}t \int w(t)f_0(t)\,\mathrm{d}t}}\,, \tag{2.2}$$

where f_0 is the reference F0 contour, $\hat{f}_0$ is its reconstruction, and w is the weighting function. The w weighting function used is given by:

$$w(t) = p(t) \cdot e(t)\,, \tag{2.3}$$

where p is the probability of voicing (POV) [66], and e is the energy computed from the utterance.

The main steps of the WCAD approach are the following[1]:

1. The approach extracts the energy contour e, the F0 contour f_0 and the POV p; based on these, the calculation of the weighting function w happens.

2. The algorithm estimates the start and the end of the phonation.

3. Next, the phrase atom is extracted by selecting it from a set of phrase atoms by maximising the WCORR with respect to F0 contour.

4. After extracting the phrase atom, local atoms are iteratively extracted from f_{diff}, using OMP with the WCORR cost function, i.e. the atom with the θ that maximises the WCORR is extracted from the set of local atoms. The local atoms are used at each iteration to update the f_{diff} and f_{recon}.

   WCAD stops local atom extraction when either: (i) f_{recon} reaches a set WCORR threshold, or (ii) the amplitude of the local atoms falls below a set amplitude threshold. The stop criterion (i) of WCAD consists in checking whether f_{recon} is below the WCORR threshold. By changing this threshold, I have a fine control over atom density, that is how many local atoms are matched. I carry out this by applying a scalar multiplier, w_{recon}, to f_{recon} when checking the fulfilment of the stop criterion. The higher w_{recon} is chosen, the more dense decomposition is obtained, i.e. a denser decomposition will consist of more accent components.

Two example atom decompositions for a Hungarian and a French utterance are shown in Figs. 2.3 and 2.4, respectively. The top plots in these figures show the F0 contours, phrase atoms and the F0 reconstructions, while the middle plots show the extracted local atoms. The bottom plots show the weighting functions.

2.5 The WCAD-based Stress Detection Approach

The individual utterances (sentences or a group of sentences together) from the database are analysed using the WCAD algorithm as outlined in Section 2.4. In order to evaluate the effect of atom density on performance, several different w_{recon} settings are used in the range of 0.025–0.5 to control the number of iterations. The extracted atoms are then grouped as

[1]The WCAD implementation is available on gitHub at `https://github.com/dipteam/wcad` (last access: 2019.10.19.). It was developed by the researchers of Ss Cyril and Methodius University of Skopje and Idiap Research Insitute.

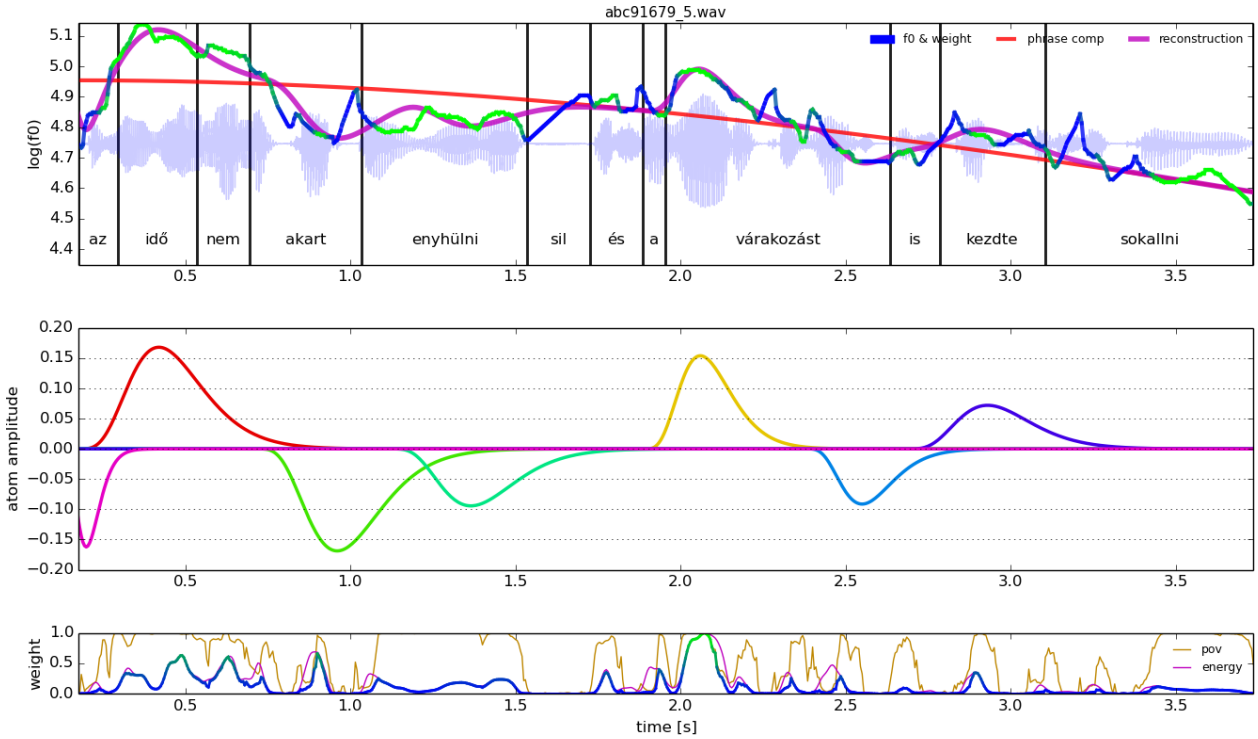

Figure 2.3: A Hungarian sentence and the WCAD output:"Az idő nem akart enyhülni, és a várakozást is kezdte sokallni. [The weather was not about to clear, and the awaiting began to be a bit much for her.]"

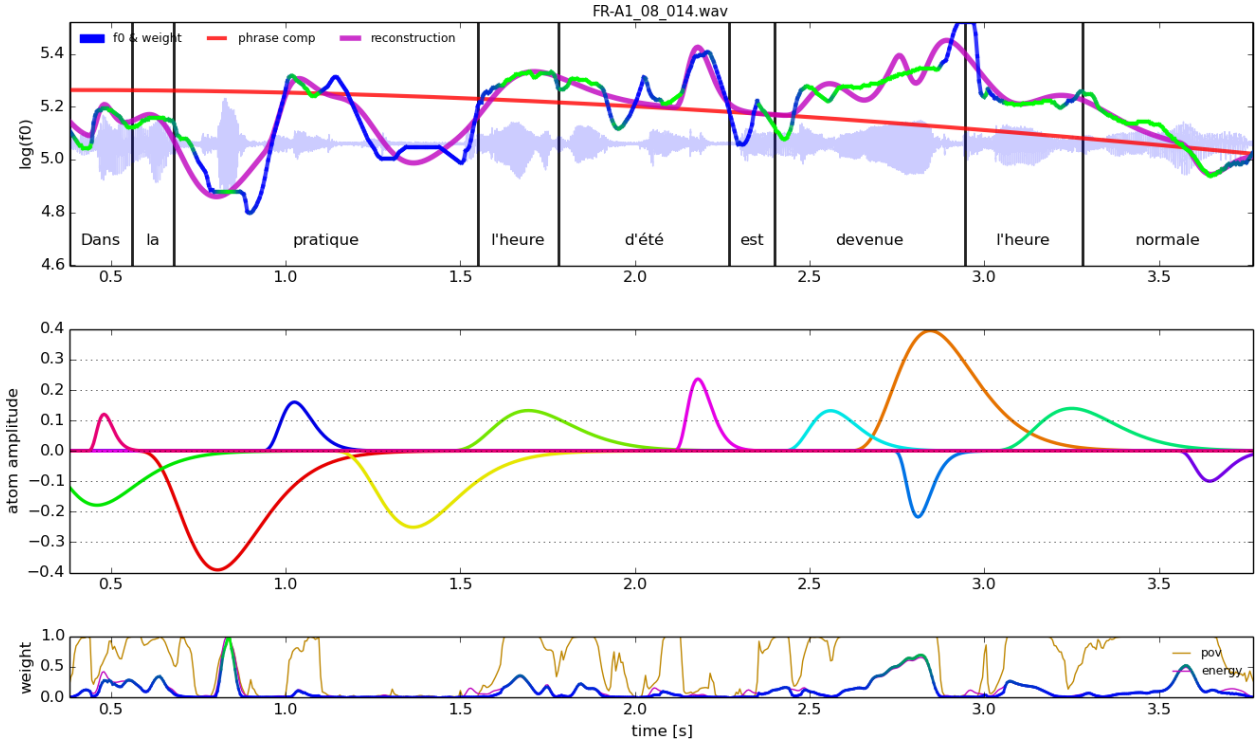

Figure 2.4: Example WCAD output for the French sentence 'Dans la pratique l'heure d'été est devenue l'heure normale. [In practice, summertime has become normal time.]"

atoms with positive (peaks) and negative amplitude (valleys), as depicted in Fig. 2.3. Their position (timestamp) is also output by the WCAD algorithm. Peaks are labelled as 'H' (referring also to a kind of high accent), and valleys 'L' (low accent)[2].

2.5.1 Mapping Atoms to Syllables

Each atom's position is mapped to the syllable within which it occurs. The syllable gets also an 'H' or an 'L' annotation notation depending on the amplitude (positive for peak and negative for valley). Syllable start and end times are derived from the phone segmentation obtained by a forced alignment (as said in Subsection 2.2). The syllabifier is a simple rule based application and it operates on the phone transcripts both for Hungarian and French. The algorithm first identifies vowels and word boundaries, then distributes inter-word consonants such that a single consonant is allowed to occur preceding a vowel (or, if two vowels are adjacent, the second one starts a new syllable). This approach yields a syllabification of sufficient accuracy for the task, especially as syllabification on phone transcripts is considered simpler and safer than on pure texts [67].

In some cases, especially in case of higher iteration numbers resulting from applying a higher w_{recon} during WCAD, syllables may be assigned having several atomic peaks or valleys (see Fig. 2.9). This happens usually at utterance initial or final positions. In such cases, the number of atoms are reduced to a single one per syllable:

- if all atoms have the same amplitude, the corresponding notation is applied (H for peaks, L for valleys);
- if the atoms have different amplitudes, a majority decision is applied (H for more peaks, L for more valleys);
- if the syllable is assigned an equal number of positive and negative amplitude atoms, the atom which appears the first is considered to be dominant.

The final atom position is the one determined by the dominant atom.

Syllables are also labelled according to their position within the word. Word-initial (WI), word-internal (or within-word, WW), word-terminal (WT) and singleton (WS) (words composed of a single syllable) are marked.

My auxiliary hypotheses regarding the correlation between atoms and syllable position are different for the two involved languages. Albeit both languages can be regarded as fixed

[2]Please note that despite employing 'H' and 'L' notations, this labelling is different from the ToBI labelling scheme.

stress, stress is located at the left edge of the PP for Hungarian [48], but at the right edge for French [47]. Some other considerations are also taken into account as explained below.

2.5.1.1 The Case of Hungarian

Given the fixed stress on the first syllable in Hungarian and other intonational characteristics outlined in [61], the auxiliary hypotheses are the following supposing that the F0 contour is approximated with WCAD:

1. H^1_{HU}: I hypothesize to find peaks on initial syllables or on singleton syllables, or in some cases on the last syllable. The latter is based on the so-called continuation rise, where the phonological (and also often the intonational) phrase ends, but the utterance continues.

2. H^2_{HU}: I expect to find valleys on terminal syllables or on singleton syllables. Indeed, this valley is equivalent to a peak on the next syllable in my approach, as I hypothesize that it signals stress on the following syllable or a following silence associated with the end of the utterance. It is due to the nature of the decomposition algorithm and the interference between the intonational and phonological phrases that stress is captured as a preceding valley here.

The H^1_{HU} auxiliary hypothesis simply says that supposing atomic peaks are stress and intonation markers, these atoms can be found on the first syllable of the words in case of prosodic stress (Hungarian is fixed stressed on the first syllable), including singleton words, or on the last syllable if a continuation rise is observed. Following this logic, if the atomic peaks are found on word internal syllables, the hypothesis is hurt.

Similarly, H^2_{HU} hypothesis says if atomic valleys are precursors of stress on the next syllable, these atoms are expected on word terminal syllables, including singleton words. As the algorithm allows a single atom per syllable, a singleton word cannot take both roles at once [3]. In Subsection 2.5.2.1, a throughout analysis will be presented for Hungarian to assess what percentage of the atomic peaks and valleys is in line with these hypotheses, and what percentage of potential word initial or terminal syllables are marked by atoms.

If this exploratory analysis confirms sufficient correlation between atoms and syllable positions, a PP detection algorithm can be built on top of them.

[3]Please note that this also means that singleton words are not counted twice (see Subsection 2.5.2.1 for further details on splitting singleton words based on the criteria whether we regard them as candidates for carrying atomic peak or rather an atomic valley).

2.5.1.2 The Case of French

French has a tendency to mark the right edges of PPs by rising F0 contour [47]. However, this is not a general phenomenon. For example, the end of an utterance or the focus may even have contours rather similar to the ones seen in Hungarian (with stress at the beginning). Observing the intonational patterns of phonological phrases in French I draw the following hypothesis, again supposing that I am approximating the F0 contour with WCAD:

1. H^1_{FR}: I hypothesize to find double markers for PPs, both at the onset and at the end, being of inverse nature. This means that either a peak at the onset and a valley at the end can be expected, or a valley at the onset and a peak at the end.

2.5.2 Correlation Assessment between Atoms and Stress

As I already mentioned in Section 2.5, the w_{recon} weight is used to control the number of iterations performed by the WCAD algorithm. By starting the experiments I observed that by $w_{recon} > 0.5$ I obtain almost identical decomposition after limiting the number of atoms per syllable to one. I can hence observe a saturation here: by setting w_{recon} higher, more iterations are run and hence more atoms are matched. After a while, new atoms typically appear almost exclusively on syllables already marked by another atom and hence are discarded as explained in Section 2.5. Thus, I conducted the stress and phrase detection experiments with $0.025 \leq w_{recon} \leq 0.5$.

An exact evaluation of stress detection can be performed if phonological phrase segmentation is addressed. With such a segmentation approach, both peaks and valleys can be simultaneously taken into account, and the evaluation is closer to possible applications as well. This is presented in Section 2.6. The following sections present some statistics for the correlation between atoms and syllable position within the words.

2.5.2.1 The Case of Hungarian

According to my hypotheses, I expect atom peaks to be mostly associated with word-initial syllables, whereas atom valleys associated with word-terminal syllables in Hungarian.

The following relative frequency-like metrics are used for Hungarian:

- the ratio of word-initial (WI) or singleton (WS) syllables associated with an atomic peak (H) vs. all syllables:

$$R1_{H,I} = \frac{c(H|syl \in WI \cup WS)}{c(H)}, \tag{2.4}$$

where $c()$ is a count operator and $\cup$ refers to set union, $|$ to auxiliary conditions and I to initial position. The higher this $R1_{H,I}$ score is seen, the more specific are atomic peaks to word-initial prosodic stress. In other words, I evaluate in which ratio atomic peaks tend to occur on first syllables (please remember that Hungarian is fixed stressed on the first syllable).

- the ratio of word-terminal (WT) or singleton (WS) syllables (referred to together as terminals T) associated with atom valleys (L) vs. all syllables:

$$R2_{L,T} = \frac{c(L|syl \in WT \cup WS)}{c(L)}. \tag{2.5}$$

Similarly to $R1_{H,I}$, the higher this $R2_{L,T}$ score is, the more specific are atomic valleys to last syllables.

- As atom peaks can be also associated with terminal syllables (hypothesized continuation rise), I calculate another score, reflecting the ratio of atom peaks which are potentially markers for a continuation rise:

$$R3 = \frac{c(H|syl \in WI \cup WS \cup WT)}{c(H)}. \tag{2.6}$$

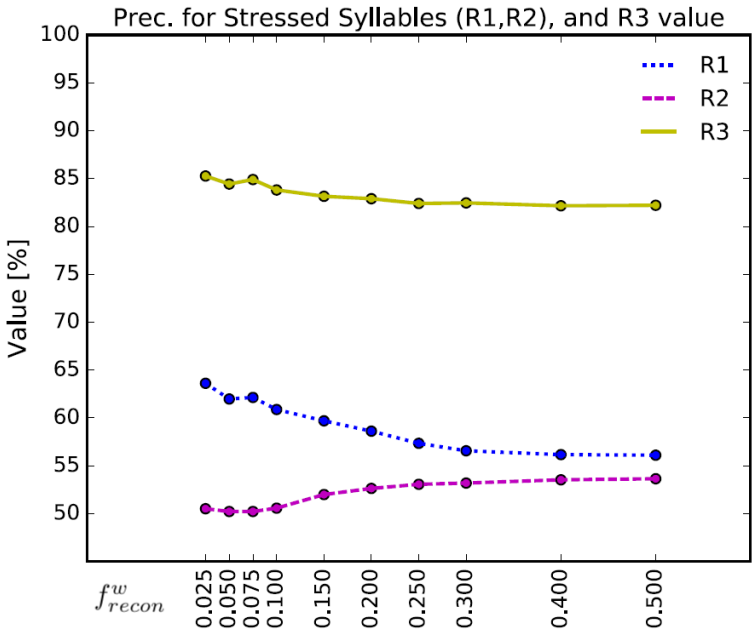

Figure 2.5: Precision-alike measures of stress/phrase boundary recovery R1, R2 and R3

Results for the R1, R2, and R3 measures for Hungarian, on the WCAD-processed data, are shown in Fig. 2.5. It can be seen that around 60% of word-initial (or singleton) syllables get marked by an atom peak (R1), and around 50% of word-terminal (or singleton) syllables are marked by atom valleys (R2). Regarding atom peaks, around 85% of them can be potentially relevant in signalling a phrase boundary (see R3).

Two more additional measures intend to reflect the ratio of peak (or valley) syllables with respect to all potential word-initial (or word-terminal) syllables, respectively. These measures are intended to give a recall alike feedback, i.e. a guess about the words potentially recoverable (the ratio of words potentially stressed). For this, singleton words shall be considered, which are word-initial and word-terminal at once. Singletons marked by either H or L are obviously classified. For the remaining non-marked singletons (singletons not receiving any atom) the mass is split between word-initial and word-terminal reflecting the distributional properties seen on the marked singletons. Word-initial and word-terminal counts are corrected accordingly. I apply the following two measures:

- the ratio of syllables with peak (H) out of all potential word-initial (WI') syllables:

$$R4_{I',H} = \frac{c(WI'|atom = H)}{c(WI')}, \tag{2.7}$$

  where $c(WI') = c(WI) + c(WS|atom = H) + c(WS|atom \notin (H \cup L)) \frac{c(WS|atom=H)}{c(WS|atom\in(H\cup L)}$.

- the ratio of syllables with valley (L) out of all potential word-terminal (WT') syllables:

$$R5_{T',L} = \frac{c(WT'|atom = L)}{c(WT')}, \tag{2.8}$$

  where $c(WT') = c(WT) + c(WS|atom = L) + c(WS|atom \notin (H \cup L)) \frac{c(WS|atom=L)}{c(WS|atom\in(H\cup L)}$.

Results for the R4 and R5 measures in Hungarian are shown in Fig. 2.6.

These results reflect recall-like measures of individual word boundaries. As not all words are stressed, the theoretical maximum for the R4 and R5 measures is well below 100%. Indeed, stressing and phrasing is speaking style and speaker specific as was outlined in [68], and approximately 33-50% of the words are stressed on average in Hungarian read speech [69]. Taking this into account, these recall rates are rather satisfactory especially as I observe furthermore that where a phrase boundary is marked by a peak, there a preceding valley is mostly missing and vice versa. This means that peaks and valleys are complementary rather than competitive regarding their boundary signalling function.

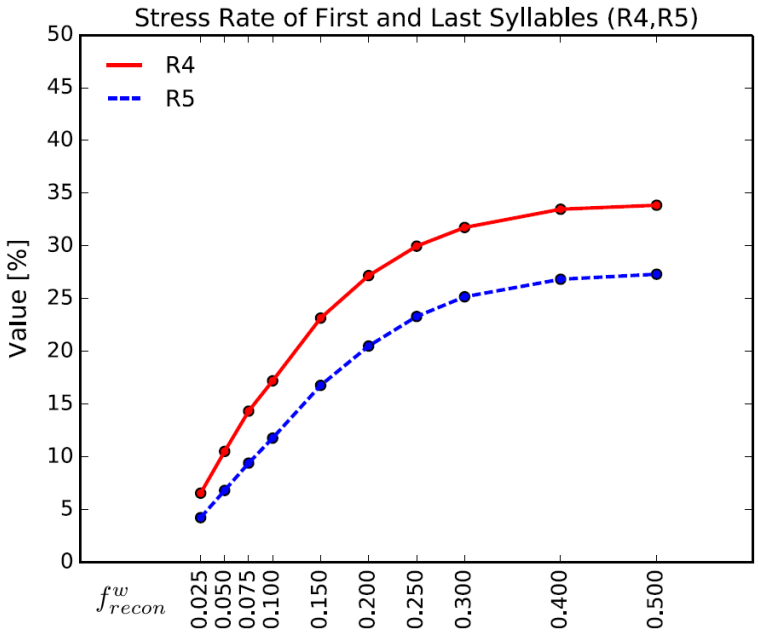

Figure 2.6: Recall-alike measures of stress/phrase boundary recovery R4 and R5

Table 2.3. shows atom/syllable and atom/word rates parametrized with w_{recon}. With $w_{recon} \geq 0.5$, atom density is quasi-permanent, a saturation occurs, therefore I present results for $0.05 \leq w_{recon} \leq 0.5$.

Table 2.3: Atom/syllable and atom/word rates with different f^w_{recon}

	w_{recon}					
	0.05	0.075	0.1	0.15	0.3	0.5
average atoms/syllable	0.16	0.22	0.27	0.35	0.47	0.5
average atoms/word	0.30	0.38	0.45	0.55	0.66	0.67
Overall atom occurrences	4617	6264	7719	10348	14457	15394

2.5.2.2 The Case of French

In French, according to my hypotheses, less characteristic correlations could be expected in terms of peaks or valleys being specific to word initial or word terminal syllables. Therefore, I move on to test hypothesis H^1_{FR} and count pairwise occurrences of peaks and valleys: where a word-initial peak follows a word terminal valley or a word initial valley follows a word terminal peak. I will refer to this special diad of two atoms as an *atom pair* seen on subsequent syllables. Please note that in this case I expect a word boundary to be marked by a co-occurrence of two atoms, which are of inverse nature.

I aggregate the occurrences of atomic pairs (including singleton words) on word boundaries (WB) vs. all occurrences of such pairs on subsequent syllables and present results in Table 2.4.

Table 2.4: The ratio of atom pairs on WB out of all atom pairs, and atoms/syllable ratios with different w_{recon}

	w_{recon}								
	0.025	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
pairs on WB/all pairs	0.8	0.77	0.73	0.66	0.66	0.67	0.66	0.64	0.63
average atoms/syllable	0.10	0.15	0.26	0.36	0.42	0.46	0.49	0.51	0.51
average atoms/word	0.15	0.24	0.40	0.52	0.59	0.63	0.65	0.66	0.67
Overall atom occurrences	467	756	1290	1741	2076	2300	2449	2598	2628

Results show 80% precision for word boundary recovery based exclusively on atom pairs at low atom density. By augmenting the WCAD threshold $f_{recon} * w_{recon}$, precision decreases as can be expected in case of denser atom layout. Nevertheless, I consider this precision value high enough for technical exploitation in prosodic stress detection and phrasing, although setting w_{recon} correctly for other corporamay be required in the future.

2.6 WCAD-based Phonological Phrasing

2.6.1 Phonological Phrasing Method for Hungarian

According to the results presented in Section 2.5.2.1., it is easy to construct a PP segmentation from atoms linked to syllables. From the phone segmentation, silence regions longer than 200 ms are preserved, then the decision is made according to the followings:

- a peak (H) signals a PP onset if it is associated with the first syllable of any word;
- a peak (H) signals a PP ending with a continuation rise if it is associated with the last syllable of any word and is followed by silence.
- a valley (L) signals a PP ending. If it is followed by silence, the utterance also terminates.

I set up these rules as hypotheses that I am going to test in the experiments.

2.6.2 Phonological Phrasing Experiments for Hungarian

Thesis I.A. [C3, J2] *I experimentally confirmed, that my atom decomposition-based phonological phrasing method significantly outperforms the HMM/GMM baseline method (by relative 7% in F1, on the investigated Hungarian corpus).*

The evaluation of PP segmentation is performed with leave-one-out cross-validation in case of the HMM system (as it is based on machine learning, most of the samples are needed for training), and directly for the WCAD-based approach. The generated PP alignment is compared to hand-labelled references. The following performance indicators are computed based on True Positive (TP), False Negative (FN) and the False Positive (FP) values:

$$\boxed{\begin{aligned} \text{Recall} &= \frac{TP}{TP+FN} \\ \text{Precision} &= \frac{TP}{TP+FP} \\ \text{F}_\beta &= (1+\beta^2)\cdot\frac{\cdot Precision \cdot Recall}{(\beta^2 \cdot Precision)+Recall}, \beta = 1, 2... \end{aligned}} \tag{2.9}$$

In my case, the number of correct phonological phrases (PPs) bounded by stressed syllables (or silent segments) are counted. Detection is regarded to be correct if the boundary is detected within a TOL vicinity of the reference timestamp. The TOL value defines hence an interval within which I accept the prediction as correct. By default, I use TOL=100 ms in my experiments, unless stated differently. Given that stress is defined for syllables, I consider 100 ms as a fair compromise, as even by fast speech tempo, the majority of syllables will not exceed TOL in duration by these settings. My rationale behind TOL is to ensure that I predict the stress on the correct syllable, as stress is bound to syllables. Obviously, higher TOL will result in better detection rates, so precision/recall plots or F-measures can be parametrized by their respective TOL value. Moreover, I calculated the average time deviation (ATD) between the detected and the reference PP boundary.

In the respective sections I have already shown that both the baseline HMM/GMM and the WCAD methods allow for the control of PP insertion "willingness" of the systems. I use these parameters ($logP_{ins}$ insertion log-likelihood in HMM/GMM and w_{recon} weighting factor in WCAD) to obtain plots of operation curves in the precision and recall space by tuning finely between dense (high recall, lower precision) and sparse (high precision, lower recall) PP segmentations.

Precision and recall plots can be seen in Fig. 2.7 for Hungarian, with the overall F_1 and F_2

scores. The overall scores were calculated as the unweighted average of F_1 and F_2 scores, at all measured operating points of each method, which equals to the different parameter values of each method (WCAD uses 10 of w_{recon}, while HMM/GMM uses 25 of $logP_{ins}$ values).

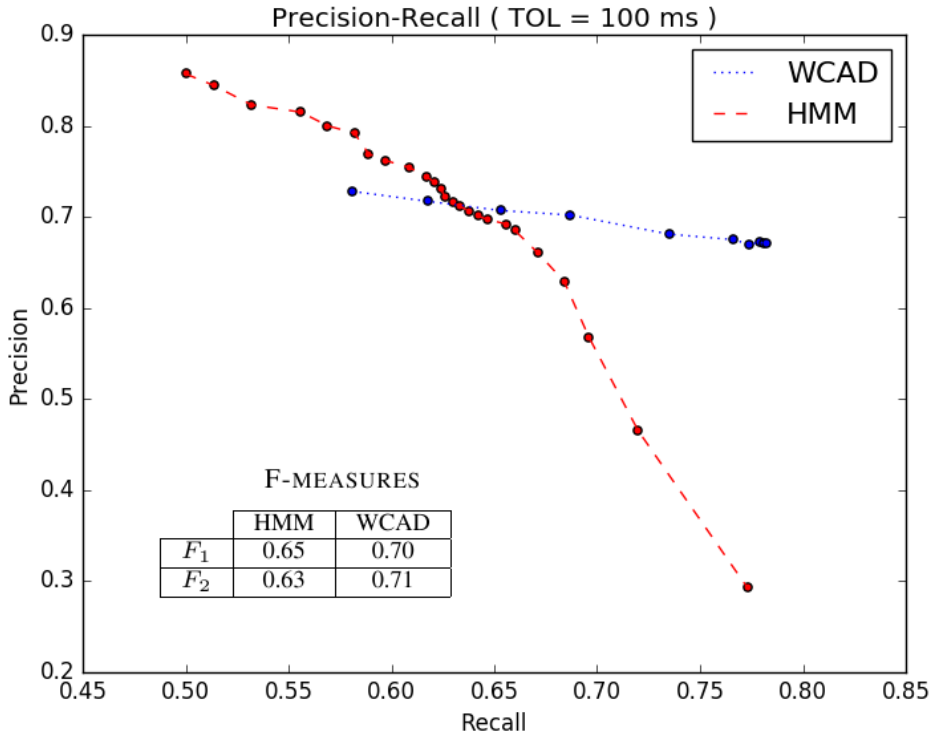

Figure 2.7: Recall, precision, and F-measures of the HMM/GMM (baseline) system and the WCAD system in phonological phrasing, Hungarian, TOL=100 ms.

The WCAD algorithm can perform significantly better than the baseline in several operating points. Which operating point is desirable for a system is determined based on task specific requirements. If high precision is required even at the expense of more type II errors (i.e. correct boundaries are missed with a sparse segmentation), then the baseline system is preferred. However, if recall should be maximized (allowing more type I errors, i.e. false boundaries are marked up with a dense segmentation), then WCAD is the better fitting choice. I measured that the performance of the WCAD approach is significantly higher than that of the baseline system in terms of F_1 by relative 7%, in overall ($p < 0.05$).

To compare the performance of the two methods on a second way, I also selected the corresponding operating points which have maximum in F_1 measure, that is, by $w_{recon} = 0.5$ in the case of WCAD, and by $logP_{ins} = -30$ in the case of HMM/GMM ($F_1 = 0.72$ and $F_1 = 0.67$, respectively); the difference is also relative 7%, in favor of WCAD method.

Concluding the previous sections on PP detection for Hungarian, results confirmed the

high correlation between atom peaks and stress and between atom valleys and upcoming stress on the next word.

Fig. 2.8 shows an example for automatic phrasing in Hungarian, for the sentence I have already used to illustrate WCAD "Az idő nem akart enyhülni, és a várakozást is kezdte sokallni. [The weather was not about to clear, and the awaiting began to be a bit much for her.]" Beneath the pane with F0 plots I provide 8 tiers of PP segmentation:

- (1) a manual reference one
- (2) a WCAD-based one by $w_{recon} = 0.1$
- (3) a WCAD-based one by $w_{recon} = 0.5$
- (4) a WCAD-based one by $w_{recon} = 1.0$
- (5) a HMM/GMM-based one by $logP_{ins} = -50$
- (6) a HMM/GMM-based one by $logP_{ins} = -20$
- (7) a HMM/GMM-based one by $logP_{ins} = 0$
- (8) a hybrid one (combined from (3) and (6), see Section 2.6.3 for more details)

As it can be seen, sparse PP segmentations tend to reveal PPs associated with stronger acoustic markers (for example a considerable rise in F0), whereas dense segmentation recovers the PP structure quite well in this case. If segmentation density is increased further, I experience that the HMM/GMM method tends to fit more PPs of lower stress (types ms and ls, as presented in Table 2.1) to places where micro-prosodic variation causes disturbance in F0, i.e. typically around plosives (leading to more false positives). The WCAD method in this case also tends to fit more atoms associated per syllable in some regions, but these extra atoms are dropped by the algorithm matching atoms to syllables.

The bottom pane of Fig. 2.8 shows the atom decomposition by $w_{recon} = 0.5$, where based on syllabification and word boundary information, the recovery of the PP structure is illustrated. Based on WCAD, it would be to some extent possible to provide identification for the PP type. If there is an H ending, that may refer to a continuation rise, whereas the strength of the stress is associated with the atom amplitude. Following silence may refer to clause ending. However, as my basic interest here was phrasing and the pure detection of stress (without classification of the strength of the stress), I did not evaluate the WCAD system from this aspect.

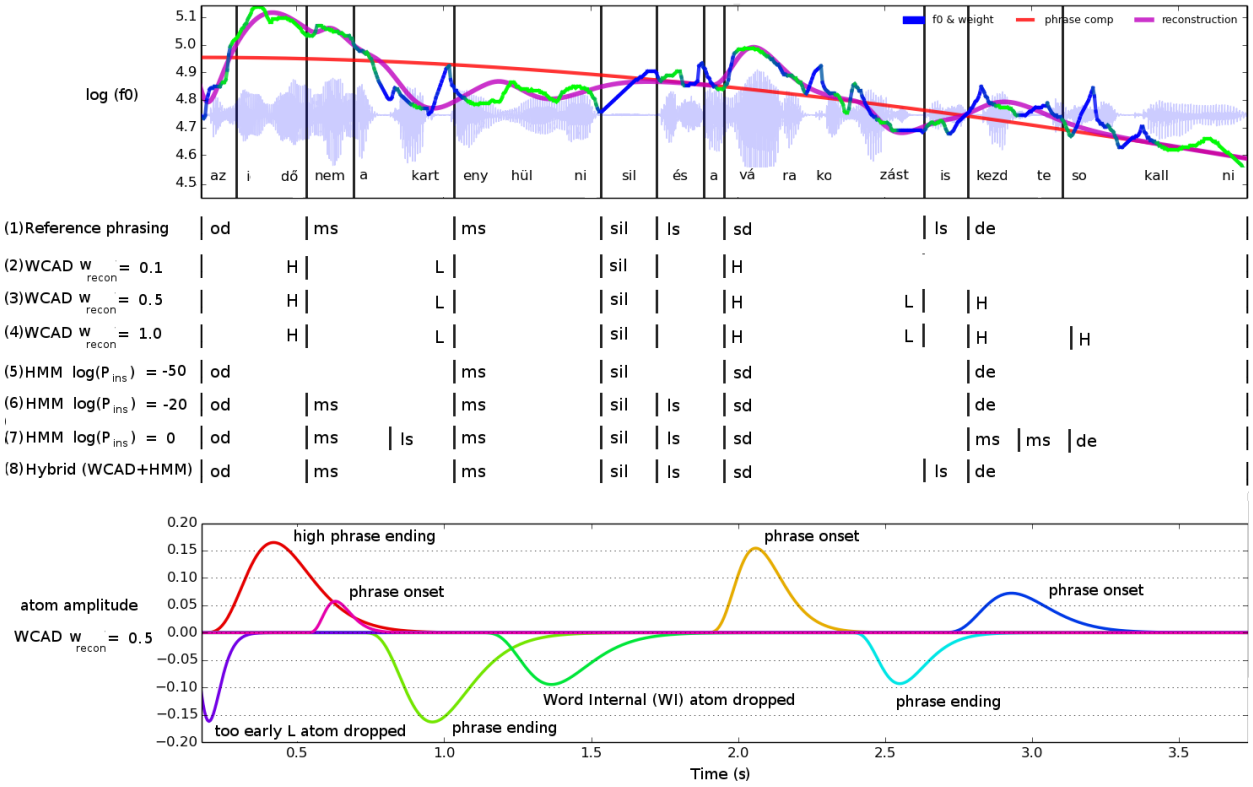

Figure 2.8: PP segmentation with WCAD and HMM/GMM by sparse and dense settings.

Fig. 2.9 shows atom decomposition of F0 with the 3 different settings for w_{recon}: 0.1, 0.5 and 1.0.

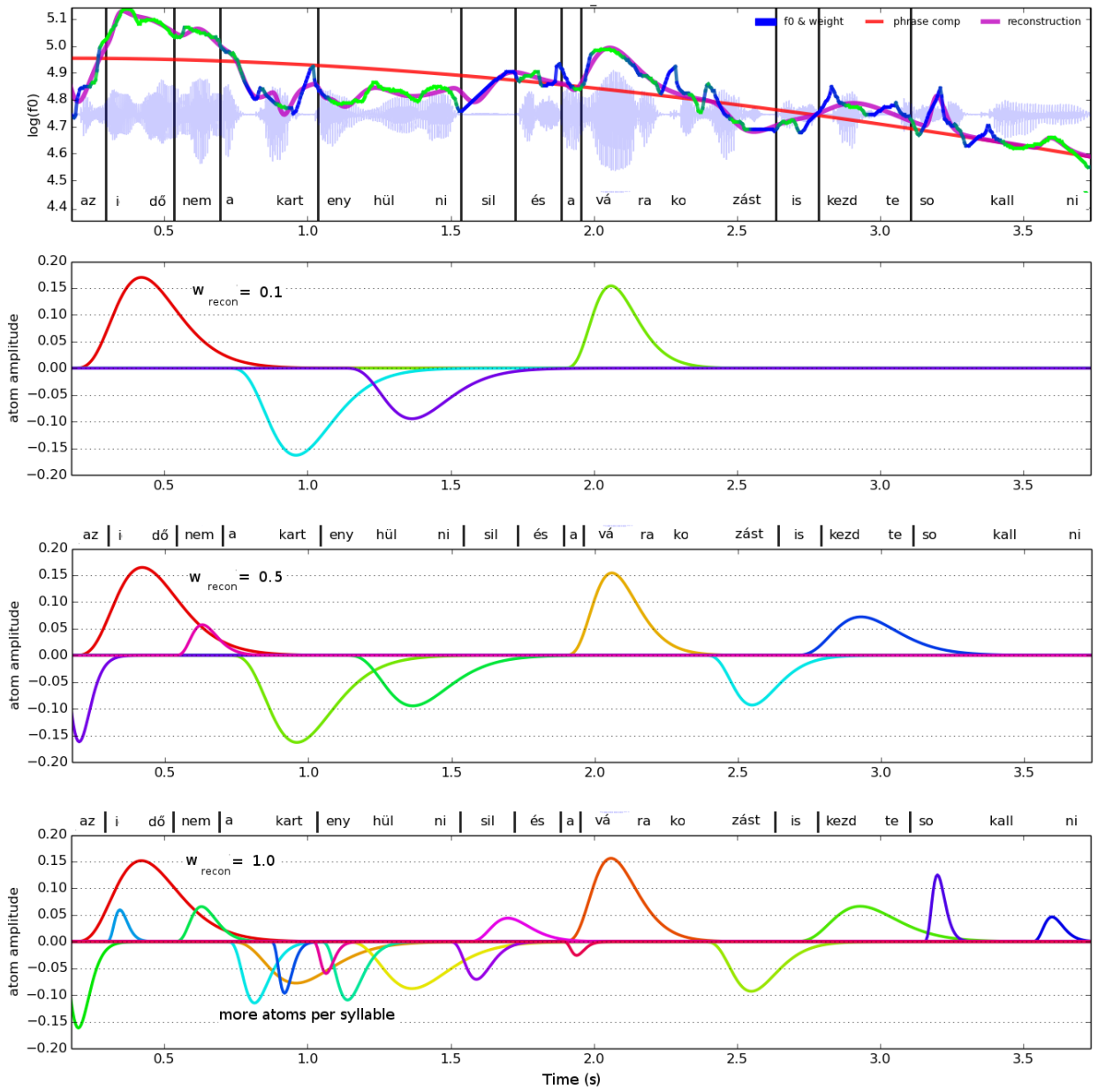

Figure 2.9: Atom decomposition of F0 with WCAD by sparse and dense settings.

2.6.3 Hybrid Phonological Phrasing Method

This section presents the hybrid approach whereby I attempt to combine the WCAD and the HMM/GMM approaches and test it on the phonological phrasing task. A straightforward scheme for a combination is to align the produced PP sequences and merge boundaries which are located close to each other, as the two systems may detect the same boundary by some time shift due to the different paradigms they rely on. For merging, I preserve boundaries further apart than 250 ms to avoid duplicate detections from the two alignments.

2.6.4 Hybrid Phonological Phrasing Experiments for Hungarian

Thesis I.B. [C2, J2] *Combining my atom decomposition-based solution and the HMM/GMM baseline approach, the obtained hybrid model yields a significant increase in the performance of automatic phrasing over the HMM/GMM baseline (by relative 11% in F1, on the investigated Hungarian corpus).*

The precision, recall and overall F-measures for the hybrid Hungarian system are shown in Fig. 2.10.

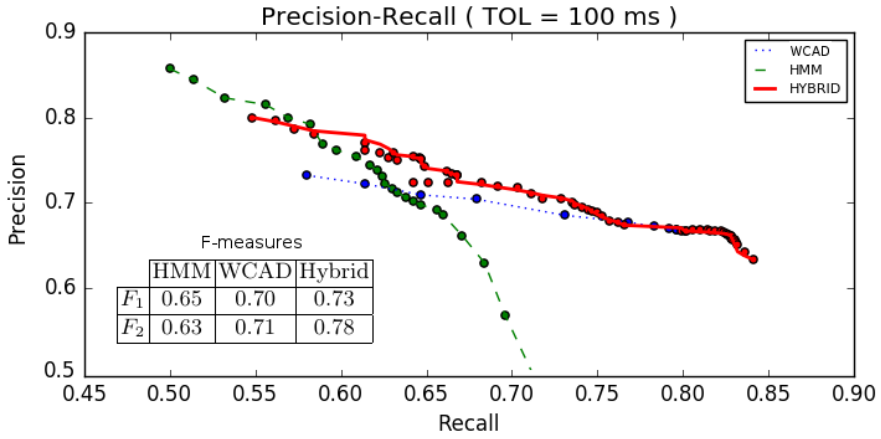

Figure 2.10: Recall and precision of the HMM/GMM (baseline), the WCAD system and the hybrid system in phonological phrasing, Hungarian, TOL=100 ms.

The number of operating points in case of hybrid model is the same as for HMM/GMM method, by overall F-score calculation. This combined approach significantly ($p < 0.05$) outperforms both individual systems for Hungarian, by 11% relative in the average F_1 score (derived from the measured operating points) compared to the baseline.

Moreover, highest F_1-score ($F_1 = 0.74$) is achieved, when the PPs of WCAD by $w_{recon} = 0.5$, and PPs of HMM/GMM by $logP_{ins} = -30$ are combined, showing relative 9.5% increase

compared to the F_1-maximum of HMM/GMM baseline. The WCAD-based systems show a slightly better time accuracy w.r.t. ATD, although WCAD-based detection knows the syllable structure, while the HMM/GMM method also finds it implicitly. Nevertheless, also the hybrid approaches preserve some of the advantage compared to the HMM/GMM baselines. By setting the tolerance TOL=50 ms, the ATD of the WCAD-based solution is still 21 ms, while the HMM/GMM solution works by ATD=31 ms. With TOL=250 ms, I obtain 38 ms with WCAD, and 58 ms with HMM/GMM solution for ATD.

2.6.5 Phonological Phrasing Method for French

According to the results presented in Section 2.5.2.2., PP segmentation for French is adopted. From the phone segmentation, silence regions longer than 200 ms are kept, then the approach checks the followings:

- a peak (H) signals a PP onset if it is associated with the first syllable of any word. In this case, the next valley (L) associated with the last syllable of a word is where the PP ends;
- a valley (L) signals a PP onset if it is associated with the first syllable of any word. In this case, the next peak (H) associated with the last syllable of a word is where the PP ends.

These hypotheses were tested during the experiments and presented in the next section.

2.6.6 Hybrid Phonological Phrasing Experiments for French

Thesis I.C. [J2] *I have experimentally confirmed, that the adaptation of my atom decomposition-based phonological phrasing method for French yields comparable results to the baseline HMM/GMM approach, while combining the two methods significantly outperforms the baseline (by relative 6% in F1, on the investigated corpus).*

For French I present results together with the hybrid system described in this section (see Fig. 2.11), as the WCAD approach did not lead to significant improvement ($p < 0.05$) over the baseline for this language. Considering the labelling accuracy with $TOL = 100\ ms$, the WCAD-based approach outperforms the baseline in some operating points, as a significant improvement is seen in average time deviation: the baseline system yields $ATD = 28\ ms$, whereas the WCAD-based approach shows $ATD = 18\ ms$. However, for French, I could

achieve a significant improvement in phrasing only for operation settings with high precision (and consequently low recall), overall the F-measures for WCAD were lower. Again, I also selected the corresponding operating points which have maximum in F_1 measure, that is, by $w_{recon} = 0.5$ in the case of WCAD, and by $logP_{ins} = -20$ in the case of HMM/GMM ($F_1 = 0.62$ and $F_1 = 0.68$, respectively); these numbers depict that WCAD approach cannot outperform the HMM/GMM in French.

I adopted the same merging strategy for French as presented in Section 2.6.3 for Hungarian. The hybrid model yields 6% relative improvement ($p < 0.05$) in the unweighted average of F_1-scores of the measured operating points, compared to the baseline. Moreover, the hybrid approach yields the best result in F_1-score ($F_1 = 0.71$), when the PPs of WCAD by $w_{recon} = 0.025$, and PPs of HMM/GMM by $logP_{ins} = 5$ are combined, reaching rel. 4% improvement compared to the baseline.

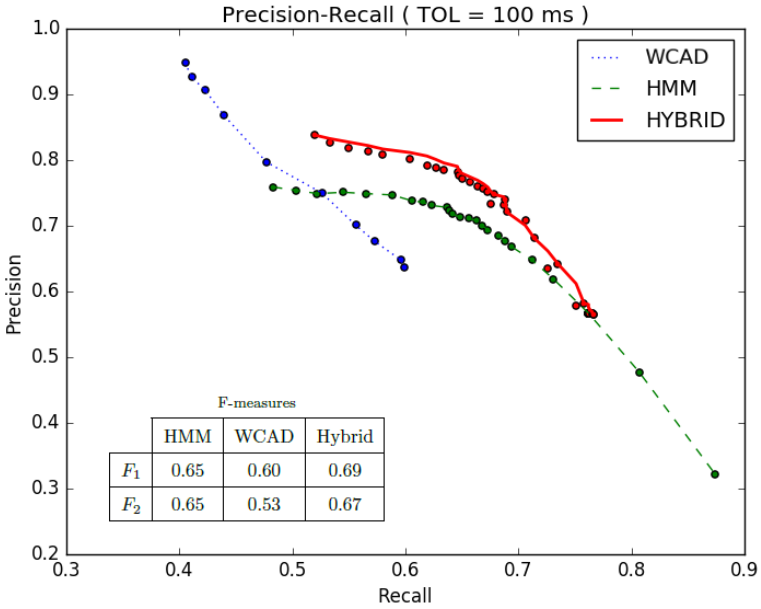

Figure 2.11: Recall, precision, and F-measures of the HMM/GMM (baseline) system and the WCAD system in phonological phrasing, French, TOL=100 ms.

In Fig. 2.12 I also present the overall results for French by applying a stricter tolerance interval TOL, limited to 50 ms only. Nevertheless, a so strict tolerance (TOL= 50 ms) interval is rarely realistic for application, as even individual phones are often longer than 50 ms. For a supra-segmental task such as phrasing or stress detection, syllables typically composed of multiple phones fit the objective better. Here I also note that if TOL is increased to 250 ms, there is no significant gain in F-measure compared to the baseline system either by WCAD alone or by the WCAD-HMM/GMM hybrid for French.

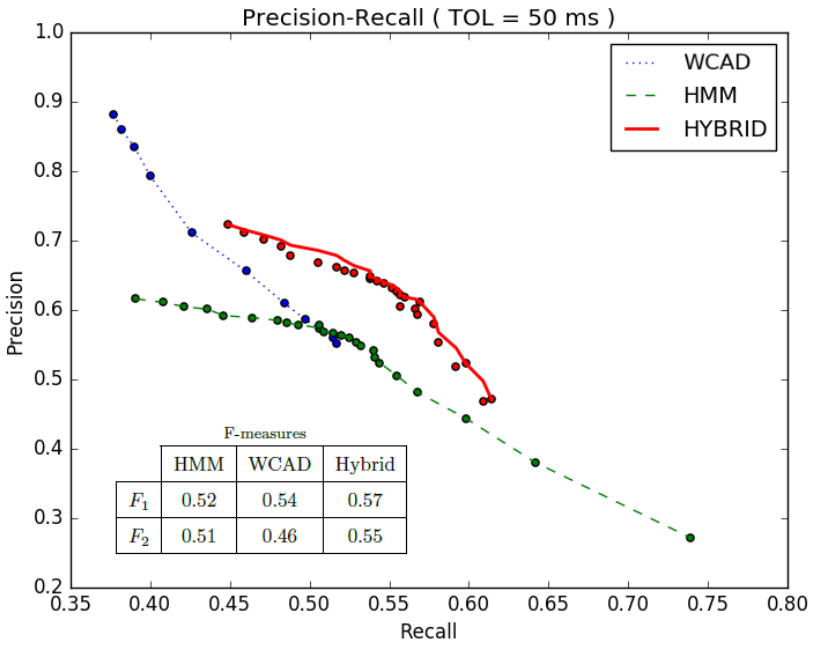

Figure 2.12: Recall, precision, and F-measures of the HMM/GMM (baseline) system and the WCAD system in phonological phrasing, French, TOL=50 ms.

Table 2.5 and Table 2.6 summarize the effect of TOL-change on the F_1-performance for Hungarian and French, respectively. It can be observed that the baseline system suffers the highest performance drop within these circumstances, as the WCAD approach is more robust in terms of its time accuracy. However, as I mentioned in the case of the analysis of ATD-values, this is not surprising, as WCAD approach knows the word alignment, whereas the HMM/GMM baseline does not.

Table 2.5: The effect of TOL-change on the F_1-performance for Hungarian

TOL - switching	**Relative decrease in performance**		
	HMM	**WCAD**	**HYBRID**
TOL = 250 ms → TOL = 100 ms	-21%	-10%	-10%
TOL = 100 ms → TOL = 50 ms	-17%	-16%	-18%

Table 2.6: The effect of TOL-change on the F_1-performance for French

TOL - switching	**Relative decrease in performance**		
	HMM	**WCAD**	**HYBRID**
TOL = 250 ms → TOL = 100 ms	-23%	-10%	-17%
TOL = 100 ms → TOL = 50 ms	-20%	-10%	-17%

Addressing differences between Hungarian and French, WCAD alone did not lead to significant improvement for French, and the overall performance gain yielded by the hybrid system was somewhat lower (but still significant at $p < 0.05$) for French than for Hungarian. I suppose that the differences may come from two sources: (i) the used databases were slightly different for the two languages on one hand; more important is, however, that I observe that the data-driven HMM/GMM approach works almost by the same F-measure for Hungarian and French. (ii) The WCAD-based algorithm however has some steps where constraints defined by rules are exploited, and data is not driving the analysis. I created these rules as explained in the respective section (Section 2.5.2), based on preliminary statistics. These were less specific for French regarding atoms signalling phrase boundaries; indeed, this is the reason why I had to use atom pairs instead of individual atoms (in contrast to Hungarian). Although, I did not made an exhaustive analysis with phonological aspects, this issue can be related to the phrase component used in WCAD, as it has a descending tendency. This fits to Hungarian phrase patterns more than to French ones, where the ascending F0 can violate this assumption. The observed pairwise occurrence of H/L or L/H atoms may be a compensation for this 'mismatch'.

2.6.7 Example for Applying Phonological Phrasing Methods for Classification

Automatic classification methods are frequently used in early diagnosis of different diseases that affect speech production. These methods can also be applied to identify speech samples from patients affected by Parkinson's disease (PD) or depressive disorder (DD). I was interested in applying automatic prosodic phrasing approaches on pathological speech samples in order to assess to what extent these tools can be useful either in characterizing in an unsupervised manner the prosodic attributes of pathological samples from individuals affected by PD and DD, or classifying samples as belonging to healthy or non-healthy individuals [4].

As my modelling level was linked to phonological phrases (PP), hence an abstract representation of the intonation and the prosodic structure was tested in general. I established that the automatically extracted PP durations and word counts can be distinctive features between the patients and the healthy group (based on statistical tests). I also analyzed the phrase distributions for additional explanations.

[4]It is a joint work with Gábor Kiss and Dávid Sztahó, as they primarily conduct researches on the detection of DD and PD. Their contribution to the pre-processing of the transcripts and the framework of support vector machine based classification is highly appreciated.

I summarize the results of [C5, C16] papers as follows: in PD speech, PPs contain fewer words, and have longer durations due to the disfluency phenomena triggered by the illness. In case of DD, the phrases got longer, but PP word counts are almost the same like the healthy subjects have. Comparing the realizations it can be seen that despite the monotony, the systems detects PP boundaries mostly at the same place (so there will be no more words in phrases on average), but the PPs characterized by a lower stress, flat intonation will have a higher proportion. These deductions are also in accordance with slight shifts observed in the overall PP distributions of the utterances showing more 'flat' speech in case of DD and impaired intonational phrase marking in case of PD.

My results show that healthy and pathological samples can be separated from each other by means of these prosodic analysers, and deep neural network or support vector machine based classifiers built on top of them. The binary classification accuracy using prosodic (PP) features was 85.6% for PD, and 78% for DD, respectively. Taking into account that the accuracy values were reached based only on these PP features, results are very promising.

Chapter 3

Automatic Punctuation with Neural Networks

3.1 Background

3.1.1 The Importance of Punctuation

In the previous sections, theses about the acoustic-driven segmentation of the speech signal were demonstrated. Punctuation recovery also can be considered as a segmentation task. As prosody is known to infer linguistic meaning and reflects the information structure [30] and modality [70] in spoken language, similar role is fulfilled by punctuation marks in written language. Punctuation marks are mainly considered as the part of the orthography in Hungarian [71, 72], only Keszler reflected on other syntactic-semantic aspects of them [73]. On one hand, the necessity of punctuation by ASR transcripts is unavoidable in case of SLU-related tasks such as parsing (including tokenization, Part-of-Speech tagging, dependency parsing, etc.), machine translation, automatic document summarization, and on the other hand, it even provides more natural and human-friendly data, i.e. punctuated closed captions for the end-users (see more in Chapter 4).

Although, proper handling and insertion of punctuation marks into the raw Automatic Speech Recognition (ASR) output received less focus in ASR technology for a long time, it became a popular topic nowadays again with the appearance of Deep Learning. We can find punctuation approaches mainly for English [74, 75, 76], but for many other languages as well; e.g. Czech [77], Estonian [75], French [78], Latvian [79], Portuguese [6], Romanian [80], Spanish [81], even Japanese [82]. For Hungarian, only the authors of [83] showed

a lightweight, pure prosody-based solution until now, while my proposed methods for automatic punctuation consider text-based aspects. Finally, I implemented text-prosody hybrid solutions as well. A review of main punctuation paradigms is presented in the next section.

3.1.2 Punctuation Paradigms

In punctuation insertion, two paradigms can be used, which are often combined: spoken and written language based approaches. When approaching from spoken language, that is audio, mostly speech prosody is considered [83, 84, 85, 86]; analysing the intonational or energy contour, together with the inter-word pause durations, the recovery of the sentence boundaries are the most effective with these approaches. Written language models for punctuation almost all exploit some sequential model and correlation of the word sequence and the punctuation marks, by usually relying on a wide context [75, 76, 84, 87, 88, 89, 90, 91, 92]. Punctuation is often combined with capitalization (also named truecasing) in a multi-task learning approach [76, 79, 80, 86, 89]. The main advantage of the prosody based models that they are more robust to ASR errors and usually require less training data than the text-based approaches, however the recently proposed written language based approaches provide a more accurate punctuation in general. The drawback of text-based models is that they are more sensitive to ASR errors and may introduce high latency due to the processing of a wide context, requiring extensive computations and involving the future context.

Early solutions demonstrated a straightforward way of providing punctuation, inspired by n-gram language models (LM). Punctuation marks, just like the individual words, show word context dependency, so incorporating them into the language model seems reasonable. This is carried out most practically by representing punctuation marks as hidden events [80, 87, 89]. Treating the punctuation problem as a sequence modelling problem lead to further innovations: finite state machine-based approaches receive a non-punctuated text as input and then they are capable of predicting punctuation as were presented in numerous works [86, 88, 90], with frameworks built on top of Hidden Markov Models (HMM), Maximum Entropy (MaxEnt) models or Conditional Random Fields (CRF), etc. These models usually allow for an easy combination of written and spoken features to obtain a hybrid punctuation model [93]. In [70], an exhaustive analysis was carried out to assess feature-wise contribution w.r.t punctuation accuracy in a MaxEnt model. The most powerful written features were the words themselves and their part-of-speech (POS) tags, whereas far the best spoken feature was the duration of inter-word pauses. Applying a monolingual translation paradigm for punctuation regarded as a sequence modelling task was also proposed in [94, 91]. The merit

of these works from my perspective is that these approaches allowed for considerably reducing time latency, which is important for ASR closed captioning.

Recently, Recurrent Neural Network (RNN)-based solutions have been proposed, which currently account for the best punctuation accuracy [75, 76, 92, 95]. Using a very large word-context mapped via word embeddings into the semantic space [10, 96], punctuation prediction becomes quite accurate, especially when coupled by an attention mechanism [75]. The authors of [97] enhanced the method of [75], applying parallel attention to prosodic features as well, including the fundamental frequency and intensity besides the "traditional" pause duration. Although these models are capable of providing good quality punctuation, their drawback is that usually they rely on large context, because the bidirectional network topologies consider not just the past but the future as well. As the lightweight punctuation solutions are desirable to run in parallel in real-time with resource demanding ASR decoders as a post-processing step, only the unidirectional approach is feasible for this case; looking at the past context only allows a low-latency solution.

Switching from the word-level investigation to exploiting character-level sequential information is motivated mostly by the idea to alleviate data sparsity problems often met at the word-level, which can be very crucial for the Hungarian language. The highly inflectional nature of this language results in several hundred thousands of different word forms. Character-level models are successfully used in several Natural Language Processing (NLP) tasks such as Part-of-Speech (PoS) tagging in historical texts [98]; language modelling [99]; Named Entity Recognition (NER) [100]. Character-level models combined with word-level models often perform the best, as reported in [101] for NER; [102] for text classification, [103] for neural machine translation; [104] for sentiment analysis etc. Pure character-level punctuation is addressed in [74], where the performance of Convolutional Neural Network (CNN) is reported to be close to the word-level baseline exploiting a Conditional Random Field (CRF) classifier.

The success of CNN in several domains [105, 106] other than artificial vision suggests that the low-level details helps both humans and artificial networks in perception and understanding, therefore an approach based on low-level – character – features is worth investigating for automatic punctuation, as it was confirmed by [74]. While the 2D-convolutional operation is valid for image processing, texts can be analyzed with the 1D-convolutional operation. It is often followed by a pooling layer (average or max) for downsampling, moreover, convolution can be the part of residual blocks as well, to boost the efficiency of feature extraction [81].

The authors of [107] compared the punctuation performance of CNN and RNN by the

transcripts of telephone calls, however, a hybrid solution was not investigated. Recently the first attempts were born combining either character- and word-level information, or RNNs and CNNs, or even more DNNs to an ensemble for punctuation restoration. Ballesteros et al. demonstrated an RNN-based language-independent approach with 5 languages, using character-based and word-based embedding together on a wide range of punctuation marks [108]; however, they missed to exploit the power of CNN architectures regarding feature extraction. The authors of [95] created an ensemble *teacher model* which consists of a Feedforward, a BiLSTM, and a BiLSTM-CRF network, yielded the best accuracy among punctuation models for English. Moreover, they applied knowledge distillation from this ensemble to a single so-called *student model*; Kullback-Leibler divergence was responsible to minimize the deviance between the *teacher model* and the smaller *student model*.

Finally, lots of Neural Machine Translation (NMT) researchers are on the party of encoder-decoder based DNN punctuation models, as a powerful alternative to RNNs and CNNs as well [109, 110, 79, 111]. Nevertheless, punctuation based on neural network models (both seq2seq and MT approaches) has been shown to outperform all other traditional approaches, such as LM, CRF, or MaxEnt-based ones.

In this chapter, I propose a lightweight word-level RNN approach compared to a MaxEnt baseline [19] for Hungarian, primarily targeted for closed captioning. This RNN model was primarily inspired by the solution proposed in [92, 75], and the successful adaptation for the Estonian language, being highly agglutinating and belonging to the same language family as Hungarian. Nevertheless, the model described in [75] is a bidirectional one relying on large future context, hence it is not directly applicable in on-line mode, which means that it brings considerable latency, that is not tolerated in real-time applications. For Estonian, prosodic features have also been leveraged; I was curious about whether only lexical features for Hungarian can lead to satisfactory results. Compared to the on-line capable [92], I added an embedding layer following the input. Attention mechanism was not involved as [75] reported slight benefit from it. As I referred to it previously, recent work on punctuation based on neural models has lead to quite complex models, but unfortunately, the focus of evaluation barely covered environment requiring on-line mode of operation, such as closed captioning, except for two works to my best knowledge [94, 91].

The closest work to mine (i.e. inserting punctuation into closed captioning for broadcast data) is presented in [94]. Instead of using an SMT-based framework, I apply neural (LSTM) models to solve the punctuation problem in an on-line closed captioning. To the best of my knowledge, my work is among the pioneering attempts to adopt neural models for on-line

mode.

Then I introduce a character-level CNN-RNN solution both for Hungarian and English. The authors of [74] compared their character-level approach to a CRF baseline with a privately-held English dataset collected for their own purpose, hence a direct comparison with my method was not possible. I analysed the performance of my CNN-RNN model against my word-level RNN approach, and I used a well-known, publicly available benchmark dataset, which consists of IWSLT TED Talk transcripts (see Section 3.5.5.1).

After that, I propose a pure textual hybrid solution for punctuation insertion, using neural networks inspired by [101]. Finally, a hybrid textual-acoustic approach is also demonstrated on a small Hungarian corpus. As Hungarian is a morphologically rich language (like Estonian) characterized with highly inflective nature and less constrained order, the extension with acoustic features is pretty important, to compensate for the variability resulting from the more variable word context. To the best of my knowledge this is a pioneering work (besides [109, 110]) in considering a character-word-prosody feature set for English as well.

The models were implemented with the Keras library [112], then trained on NVIDIA GPUs, as they are a better choice for performing mathematical matrix operations (even in parallel) than CPUs.

3.1.2.1 Deep Feedforward Neural Networks

The first DNN approaches called deep feedforward neural networks (DFNN) or multi-layer perceptrons (MLP) are discussed, as the basis of more complex architectures (like Convolutional or Recurrent Neural Networks (CNNs, RNNs)). The illustration of a simple neuron is shown in Figure 3.1; considering its main tasks, the summation and activation (which depends on the value of weighted sum of the input features). Nowadays non-linear mathematical functions are used as activation functions, such as sigmoid, tanh, or rectified linear units [113].

As one neuron is not sufficient to solve complex (e.g. multi-class) classification tasks, the neurons are connected to each other, organized into a neural network. Figure 3.2 shows the schema of an MLP network (ignoring *bias*). In this architecture, usually more hidden layers are involved, containing numerous (some hundred or thousand) neurons, hence the *deep* name comes from, and the information flow is unidirectional between the neurons, that is why it is called $feedforward$.

These networks can act as universal function approximators [114], where the weights of the network are the parameters of the function. The goal is to learn connections between

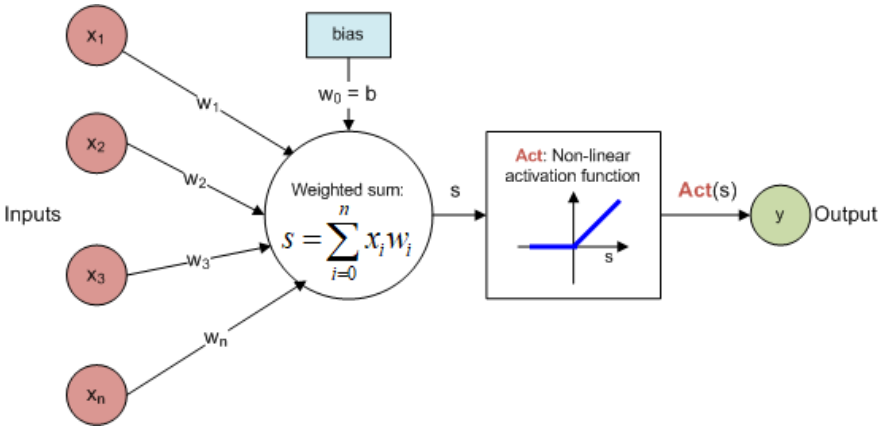

Figure 3.1: Structure of a neuron

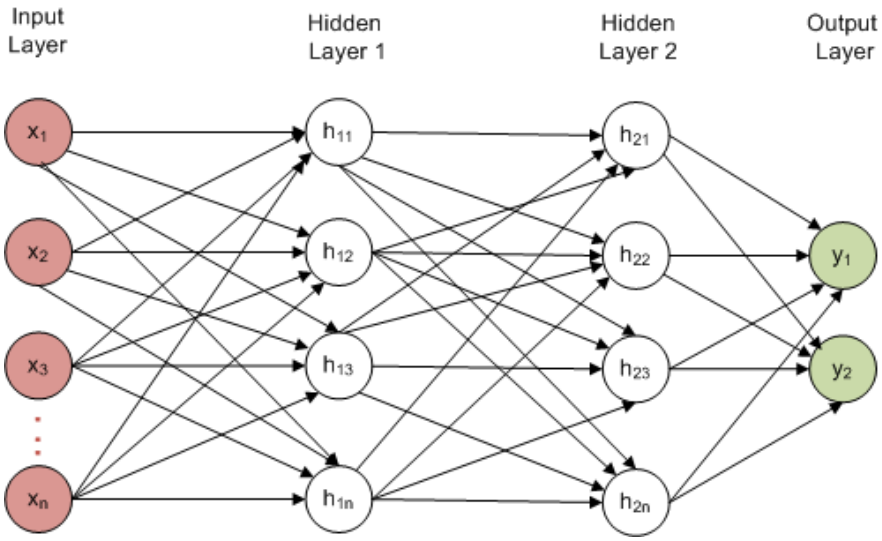

Figure 3.2: Structure of a feedforward neural network

given (X, Y) input-output pairs, with the continuous adjustment of these weights, minimizing the cost (or loss) function, which is derived from the deviation of the target output and the actual output. In the beginning, these weights can be randomly initialized [115].

During backpropagation [116] used to train neural networks, the calculated error is traced back from the final layer in the reverse direction. The adjustments happen in the opposite direction compared to the gradient of the loss function, and the *learning rate* controls the extent of it. A common approach is when *stochastic gradient descent* technique with a mini-batch concept is applied, an average loss is always calculated on a held-out subset of training samples or on a separate validation set. In the ideal case, this loss is iteratively decreasing during the training, and the model converges. The iterations correspond to epochs; in one epoch, the entire training set is processed. There are several optimization techniques [117],

which are applied to calculate the learning rate and/or the loss (including adding regularization terms), to ensure the stability of the training. Different methods (called often simply optimizers), such as e.g. RMSProp [118], Adam [119], etc. adjust the learning rate and calculate the loss slightly differently, but the main goal remains to ensure true convergence (i.e. not to get locked into a local minimum) and optimizing the speed of this convergence. Finally, after the training – which includes iterative evaluation on the validation set – the model is evaluated typically on a separate test set.

The DFNNs are hence capable of mapping input vectors to output vectors, but they handle the inputs independently and miss to catch their contextual information. This gap is eliminated by Recurrent Neural Network architectures, presented in the next section.

3.1.2.2 Recurrent Neural Networks for Sequence Labelling

Recurrent neural networks (RNNs) are often used in a sequence-to-sequence (seq2seq) modelling approach, where the corresponding task is referred to as neural sequence labelling. In the case of sequence tagging, the neural network receives a series of inputs and assigns an output (label) to each input. RNNs utilize neurons/cells which have a memory unit, preserving some information about past states of the cell. The memory unit itself is regulated by the data flow, as well as the output of the cell, which is combined from a weighted contribution of the current input and the memory unit. The regulating weights are learned during the training phase. Another technique providing a regularizing effect is Dropout [120]; dropping out randomly chosen hidden (and visible) units in a neural network in a mini-batch can help to avoid of overfitting. The first variants of RNN models suffered from the exploding and vanishing gradient problem [121] during the long-term dependencies of the sequences, causing insufficient learning.

These problems were alleviated with the mentioned of Long-Short Term Memory (LSTM) [122] and GRU [123] cells, which have inner memory. The key factor is the "gating" mechanism. As Figure 3.3 shows, LSTM has three gates (forget, input, output), while GRU has only two (reset and update), to maintain the memory of the cell.

My RNN models for automatic punctuation are built up from recurrent layers with LSTM cells, as connecting them sequentially leads to powerful sequential models [124, 125, 126]. For the punctuation task, the data is modelled as a time series, in a synchronized *many-to-many* approach; typically each LSTM cell receives input features at a given time frame of a sequence, and an output label is generated for each input. The output label can be a punctuation mark (e.g. comma), or a blank label, meaning that no punctuation mark is

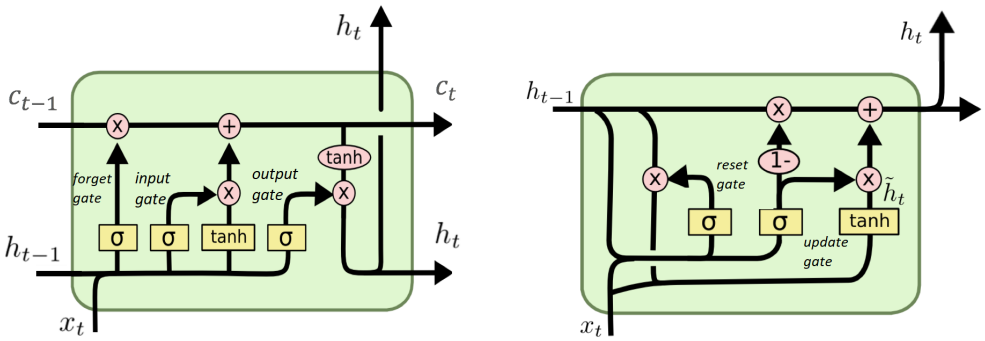

Figure 3.3: Structure of an LSTM (left) and a GRU cell (right)[1]

predicted to a given input in the sequence. The capability of unidirectional LSTM solution is to use only the past context besides the current time frame. However, for a sequence labelling task, it is more common to incorporate future features into the processing framework, that is, the output of the network at time t depends on inputs ranging from $t-k..t..t+k$. This is usually more effective if the bidirectional (from past to future and from future to past) flow of the information is allowed within the network (e.g. Bidirectional LSTM, BiLSTM [127]). Obviously, the future is not known, so technically such networks wait until future samples are available (for example, until final words of a sentence appear), and delay their output accordingly. I will refer to this setup as *off-line mode*. Operation by limited future context will be referred to as *on-line mode*. I experienced during my experiments, that limiting the future for a single word is more efficient for punctuation task. Depending on the length of the word and the speech rate, this results in a latency of several hundred milliseconds in the worst case when using spoken input.

Adapting this framework to word sequences I consider a token at each time frame, that is word by word by indexing the words by t. After that, I used this methodology for character-sequences and prosodic feature vectors as well. Considering the word input features, instead of using the surface form of the words, the involvement of low-dimensional word embeddings into recurrent neural networks is popular nowadays in case of other NLP-tasks than punctuation as well [128, 129].

[1]Based on source: https://github.com/roomylee/rnn-text-classification-tf (last access: 2019.10.19.)

3.1.2.3 The Importance of Word and Character Embeddings

In a continuous vector space, the vector distance between two words is related to their syntactic-semantic relationship; this vector representations are also called word embeddings [10, 130, 131]. Learning distributed representation of the words are mainly originated from neural language modelling, capturing many of the word neighbourhood characteristics from a huge text corpus (like Google or Wikipedia texts) [132]. Albeit, more and more ideas come up nowadays for distributed word representations, I use pre-trained embeddings derived from two popular approaches, hence I present just these techniques here.

Word2vec [10] offers DFNN-based methods with unsupervised learning, utilizing contextual information in two ways. In the Continuous Bag-of-Words (CBOW) approach, the goal is to predict a target word based on its surrounding context (the size of this context can be configured during the training procedure). In the opposite, in Skip-gram approach the context is predicted for a central target word. The real-valued word vectors are extracted from a hidden layer of a feedforward neural network. Figure 3.4 shows an illustration for the proximity of other countries close to Hungary in a Word2vec space.

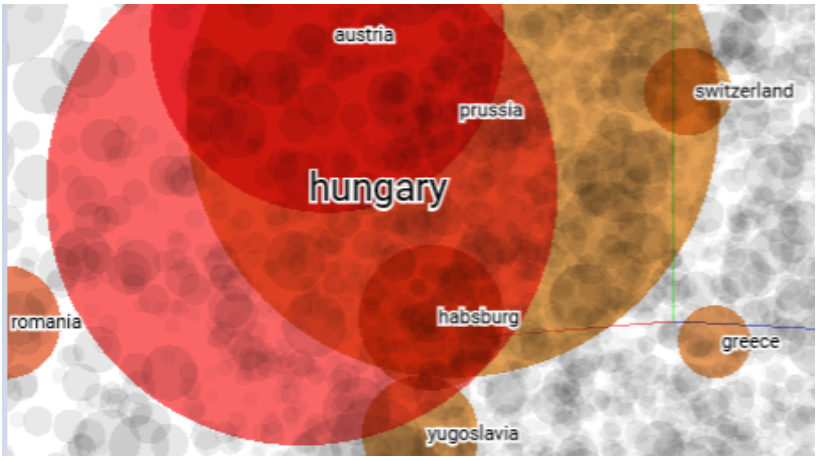

Figure 3.4: The *neighbours* of Hungary in a Word2Vec representation[2]

The other word embedding approach is called GloVe [96], where the basic idea is to use global count statistics, calculating the embedding weights based on a word-word co-occurrence matrix, extending the local context window method of word2vec. The authors also proposed a weighting formula to emphasize the meaningful co-occurrences.

However, while word embeddings act as the fundamental base for many NLP tasks in English, Out-of-Vocabulary issues are often encountered, especially for the morphologically

[2](edited via https://projector.tensorflow.org (last access: 2019.10.19.)

rich languages like Hungarian (e.g. the plural form of a noun can be already OoV), as there are many words only with rare occurrences, which can impair the neural network learning. To alleviate this problem, character embedding-based solutions became popular and successfully applied, despite of losing the much meaningful relationships, which were applicable between words [99, 133, 134]. Moreover, the Convolutional Neural Networks have important contribution to this success; these models are presented in the next section.

3.1.2.4 Convolutional Neural Networks for Low-level Text Features

The success story of Convolution Neural Networks (CNNs) began with the topic of artificial vision, as it counts as a powerful pattern recognition approach [135, 136]. As the way of image processing is hierarchical in our brain, likewise, the deep CNNs with numerous hidden layers are able to detect different level of features from the pixels. First, the local connections are learned (for example the edges of a face are found), then even bigger parts, like eye or lips are recognized, and finally, a whole face as a complex representation is captured [7].

CNNs are built from special neurons, which are acting like a filter or feature detector. They process only a little part of an image (or it can be other signal, e.g. speech or text), but they do it many times: The whole input will be processed by the different convolution filters, owing to a sliding window concept (with a configurable stride). Finally, different feature maps are generated. For instance, the convolution filters move in a grid in case of pictures (left-to-right and/or top-down), and their weights are learned automatically during the training.

After the convolution operation, a pooling step is performed, which is basically a down-sampling; a specific region of feature maps will be represented with one neuron, calculated from the average of output values, or just simply selecting the maximum, keeping the most salient information [137]. Again, a sliding window is used for this purpose. After the pooling operation, usually fully connected layers are stacked.

While the benefit from 2D-convolution operation is quite obvious for image processing, the characteristics of local connections and compositional aspect hardly can be aligned with text inputs. After the pioneering work of Collobert et al. [138], the real advent of 1D CNN modelling for text-related classification tasks started in the mid 2010's [139, 140].

For the text inputs, only sequential processing (left-to-right filtering) is applicable. Usually 1D (temporal) convolution is performed on sentence-level through word embeddings, but nowadays it can be applied to character / character-sequence inputs as well. After the pooling layer, in my sequence labelling approach, a BiLSTM layer follows, to capture

possible long-range dependent information from the patterns extracted from character (embedding) n-grams. Please note, that the fixed-length sequences (chunks) do not correspond to sentences, especially when they contain 150-200 words.

3.1.2.5 The Baseline MaxEnt Model

Maximum Entropy (MaxEnt) models were first suggested for POS Tagging [141], where each sentence was tokenized into words and described as a sequence. Tokens have a set of associated features, and as a supervised learning solution, the output labels are assigned to the token series. To determine the set of features, the MaxEnt model defines a joint distribution over the available tags and the current context. The context can be controlled with a radius parameter.

I use the MaxEnt model only with word-related input features; punctuation prediction is obtained by the Huntag [19] open-source[3] Maximum Entropy Markov Model-based Sequential tagger, which by design, is a language- and task-independent framework. The same train, validation and test subsets were involved in the experiments of MaxEnt models and RNN models to ensure fair comparability (see Section 3.2.1 for the Hungarian dataset and Section 3.5.5.1 for the English dataset). Words are all converted to lowercase (hence no capitalization information is exploited).

The radius parameter of the MaxEnt tagger determines the size of the context considered before predicting a punctuation mark. By default, left (past) and right (future) context is taken into account. I will refer to this setup as *off-line mode* as similarly to the case of the RNN approach. I limited the future context also to a single word, referred to as *on-line mode*. In the experiments I use round brackets to specify left and right context, respectively. Hence (5,1) means that I consider 5 past and 1 future tokens actually.

[3]The source code is available here: https://github.com/recski/HunTag (last access: 2019.10.19.)

3.2 The Word-based RNN Model

In the data pre-processing step, the data is split into short, fixed-length sub-sequences (chunks) with no overlap and no skip, so each word token is mapped exactly once. The length of chunks is a free parameter to be optimized (see Table 3.1). A vocabulary with $k+1$ entries is built from the k-most common words in the training set plus an *Unknown* entry which is used to map rare words to a common vector. Additionally, an *EOS* entry indicates the end of the subsequence. Incomplete chunks were padded with zeros to fit the fixed length. An embedding weight matrix was added based on pre-trained embeddings for the tokens in the vocabulary.

Fig. 3.5 shows the networks used during the experiments. The fixed length chunks are input to the model. Words are indexed by $t = 1..T$, where T equals to the length of chunks. Each word in the chunk is mapped to a vector representation using the embedding matrix, so the preprocessed sequences are projected into the embedding space. In Fig. 3.5, x_t represents the word vector x for word t (t^{th} word in the chunk, regarded as a sequence). The following layer is composed of LSTM or BiLSTM hidden cells. This structure altogether captures a context for x_t. The output is obtained by applying a *softmax* activation function to predict the y_t punctuation label for the slot preceding the current word x_t.

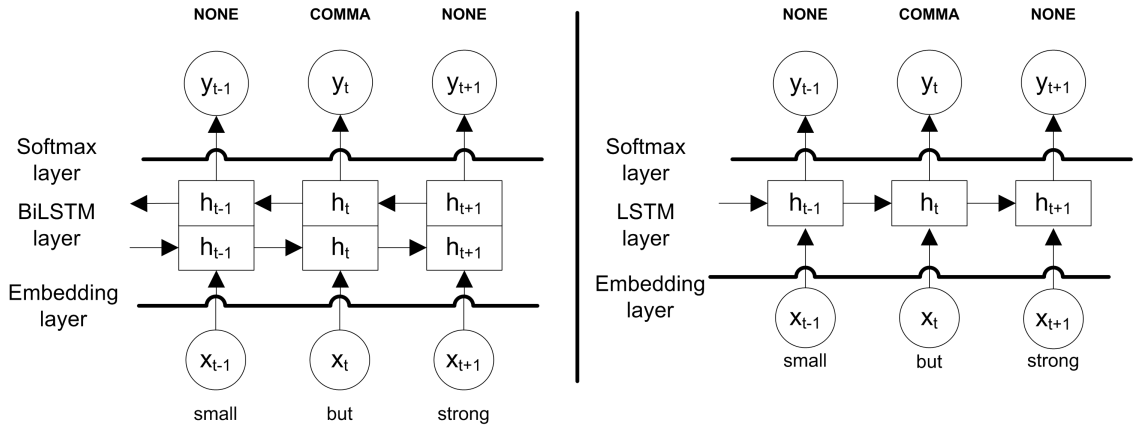

Figure 3.5: Structure of WE-BiLSTM (left) and WE-LSTM (right) RNN model

The Hungarian word-based RNN punctuation models were trained on the 100K most frequent words in the training corpus, by mapping the remaining outlier words to a shared *"Unknown"* symbol. The models use 600-dimensional pre-trained Hungarian word embeddings [142]. These word embeddings were trained with CBOW approach of word2vec (see Section 3.1.2.3), including cca. two million word entries altogether, derived from the Hungarian National Corpus [143] and from the Hungarian Webcorpus [144]. I assume that the

relative high dimensionality of the embeddings comes from the highly agglutinating nature of Hungarian. During training, I use categorical cross-entropy cost function for the softmax function, as automatic punctuation is considered a multi-label classification task. I also let the imported embeddings to learn, because they can adapt to the task this way.

A systematic grid search optimization was performed for hyperparameters of the RNNs on the validation set: length of chunks, vocabulary size, number of hidden states, mini-batch size, optimizers. I also use early stopping to prevent overfitting, controlled with a *patience* variable. Table 3.1 summarizes the final values of each hyperparameter used in Hungarian WE-BiLSTM and WE-LSTM models (WE refers to Word Embedding).

Table 3.1: Hyperparameters of WE-BiLSTM and WE-LSTM models

Language	**Model**	**Chunk Length (#words)**	**Vocab. Size (#words)**	**Word Embedding dimension**	**#Hidden states**	**Batch size**	**Optimizer**	**Patience**
HUN	WE-BiLSTM	200	100 000	600	512	128	RMSProp	3
HUN	WE-LSTM				256			2

As for the MaxEnt setup, I also use a low latency and lightweight on-line mode (WE-LSTM), and a robust off-line mode exploiting the future context (WE-BiLSTM).

3.2.1 The Hungarian Broadcast Dataset

The Hungarian Broadcast Dataset is derived from public broadcasts of the Media Service Support and Asset Management Fund (MTVA), Hungary. The raw data contains various TV genres such as weather forecasts, broadcast (BC) news and conversations, magazines, sport news and sport magazines. The dataset used for the training is a subset with manual transcription including manual punctuation.

The punctuation marks addressed in the experiments include commas, periods, question marks and exclamation marks, as these are regarded to be primarily important in terms of readability. The colons and semicolons were all mapped to commas, whereas all other punctuation marks were simply removed from the corpus. The dataset is split into non-overlapping train, validation and test subsets (I reserve a disjunct 20% of the training corpus for validation). Characteristics of the used subsets are summarized in Table 3.2.

I used both manual and ASR transcripts in order to assess punctuation performance. ASR transcripts correspond to the case where the system is used in real-time operation mode, manual transcripts are used for supervised training and benchmarking.

As the manual transcripts of the Hungarian Broadcast Dataset (including the test subset) contain manually inserted punctuation marks, the training and evaluation can be done on

Table 3.2: Statistics of the Hungarian Broadcast Dataset

Genres	Training & Validation					Test					
	#Words	#Com	#Per	#Que	#Excl	#Words	#Com	#Per	#Ques	#Excl	WER
Weather	478K	40K	31.5K	30	730	2.4K	250	200	0	20	6.8
Brc.-News	3493K	279K	223K	3.5K	4.6K	17K	1.5K	1K	20	50	10.1
Sport news	671K	55K	39.5K	280	2K	6K	500	400	2	30	21.4
Brc.-Conv.	4161K	533K	225K	26.5K	4K	46.8K	6.3K	2.6K	250	130	24.7
Sport mag.	-	-	-	-	-	22.7K	2K	1.4K	100	50	30.3
Magazine	4909K	732K	376K	72K	36K	10.4K	1.5K	700	150	70	38.7
Mixed	1526K	187K	102K	11K	11.4K	30.7	4K	1.7K	280	150	-
ALL	**15238K**	**1826K**	**997K**	**113K**	**58.8K**	**136K**	**16K**	**8K**	**800**	**500**	**24.2**

transcripts with similar quality. In ideal case, the ASR transcripts should also contain manually inserted punctuation marks for training and testing purposes as well. However, in the most common case the ASR transcripts do not contain target labels (i.e. punctuation marks); it is also true for the Hungarian Broadcast Dataset. The manual punctuation of these texts is not done because it is time-consuming. Moreover, it is also challenging for linguistic expert to put punctuation marks into an altered word chain.

As a consequence, there is a mismatch, as the ASR transcripts (which is used only for testing) are evaluated after training with manual transcripts, but there are also other issues with this set-up. The manual reference may not be the ideal basis for training and assessing punctuation performance for ASR transcripts, as the altered word chain may require a different punctuation pattern if such a pattern can be meaningful at all.

Moreover, the only solution for the evaluation is to perform a comparative assessment between the reference and ASR transcripts with an alignment algorithm which considers Levenshtein-distance (it is originally used for ASR evaluation). In my case, the punctuated manual (reference) set is aligned to the punctuated ASR set, where the punctuation marks are handled as separate word tokens.

To sum up, I use a common training set with manual punctuation, then the automatically punctuated Hungarian ASR transcripts are evaluated based on the result of an alignment with the manually punctuated reference transcript. Please note that in case of English experiments, I used a fixed ASR test set (pre-processed and verified by the author of [145]), so there is no need to perform the alignment task.

Hereby I show an example for the result of an alignment between the two transcripts. The output shows insertion errors (which can be an extra word or punctuation mark in the ASR hypothesis), deletion errors (i.e. missing punctuation label or word), and substitution errors (green: reference, red: wrong word or punctuation label), besides the correct tokens.

<hirado1.txt>
[...]

insertion	egy		<word insertion error, not counted in the evaluation>
correct	kolozsváron		
correct	a		
correct	demokrácia		
correct	önkéntesei		
correct	segítenek		
correct	a		
correct	nyomtatványok		
correct	pontos		
correct	kitöltésében		
correct	*fullstop*		<correct punctuation mark, counted in the evaluation>
substitution	ha	a	<word substitution error, not counted in the evaluation>
correct	kell		
deletion	*comma*		<punctuation deletion error, counted in the evaluation>
correct	lépésről		
correct	lépésre		
correct	elmagyarázzák		
correct	*comma*		<correct punctuation mark, counted in the evaluation>
correct	hogy		
correct	melyik		
correct	borítékba		
correct	mit		
correct	kell		
correct	tenni		
deletion	*fullstop*		<punctuation deletion error, counted in the evaluation>
substitution	itt	a	<word substitution error, not counted in the evaluation>

[...]

Based on this output, the performance indicator of the punctuation restoration can be calculated (ignoring the word errors).

3.2.2 The Hungarian ASR System

The ASR used for the experiments is the closed captioning system presented in [12]. Both the acoustic and language models are optimized for the closed captioning of broadcast data representing the covered six genres (broadcast news and conversation; sport news and magazines; magazines; weather forecasts). The language model for the ASR was trained on the corpus used for the training of the punctuation model with the SRILM toolkit [146]. The deep neural network based acoustic models were trained on 500+ hours of transcribed speech using the Kaldi ASR toolkit [147].

The overall word error rate (WER) on the entire punctuation test set was 24%, albeit WER showed a large variance depending on genre (see Table 3.2). Please note that the WERs of Hungarian and English ASR systems are not directly comparable. As it was mentioned in Section 3.1.1, due to the highly inflective nature of the language, a recognition error in a prefix or a suffix can make a whole word incorrect, hence WER tends to be higher for Hungarian than for English tasks, even if the subjective quality measures of the ASR produced automatic transcriptions are almost the same [148]. Moreover, compound words output as two separate, but correctly recognized words account for 2 word errors.

3.2.3 Experimental Results for Word-based Punctuation

Thesis II.A. [C6, C7, C17] *I have experimentally confirmed, that my word-based RNN model for automatic punctuation significantly outperforms the Maximum Entropy-based baseline system, in off-line operation mode, for Hungarian language (20% relative improvement on reference transcripts, and 10% relative improvement on ASR transcripts in SER).*

This section presents the word-based punctuation recovery results for Hungarian. For the objective evaluation, I use again standard information retrieval metrics, such as Precision (Pr), Recall (Rc), and the F1-Score (F1) (described earlier in Equation 2.9). In addition, I also calculate the Slot Error Rate (SER) [149], as it incorporates all types of punctuation errors – insertions (Ins), substitutions (Subs) and deletions (Dels) – into a single measure, and weights them equally:

$$SER = \frac{C(Ins) + C(Subs) + C(Del)}{C(slots_p)}, \tag{3.1}$$

where $slots_p$ refers to the number of punctuation marks in the reference, and $C(.)$ is the count operator. Unlike the F1-score, with lower SER values model performance is higher.

I compare the performance of my RNN-based punctuation recovery system (see Subsection 3.2) to the baseline MaxEnt sequence tagger (see Subsection 3.1.2.5) on the Hungarian broadcast dataset. Both approaches are presented in two configurations. While in the *off-line mode*, the future word context is also exploited to achieve the best result with the given features and architecture, I also do a performance analysis when switching to *on-line mode*, where the punctuation restoration depends mainly on the past word context (I allowed future context involving only a single word). This configuration is intended to fit to real-time closed captioning requirements with low latency. In both modes, punctuations are predicted for the slot preceding the target word (at t) in the input sequence. The hyperparameters of all presented approaches and their configurations were optimized on the validation set as explained earlier (see Section 3.2).

The obtained results are presented in Table 3.3 for the manual transcripts and in Table 3.4 for the automatic (ASR) transcripts, respectively. To recall, in the notation of MaxEnt models (i, j), i stands for the backward (past), whereas j stands for the forward (future) radius.

Table 3.3: Punctuation restoration results for Hungarian manual transcripts

Manual Transcript	**Model**	**Comma**			**Period**			**Question**			**Exclamation**			**SER**
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(19,19)	72.5	59.6	65.5	52.1	40.0	45.2	**55.7**	21.8	31.3	31.1	31.5	31.3	63.5
	WE-BiLSTM	**72.9**	**71.2**	**72.0**	**59.1**	**56.1**	**57.6**	52.4	**38.7**	**44.5**	**51.3**	**36.1**	**42.4**	**50.1**
On-line mode	MaxEnt-(25,1)	71.8	58.1	64.2	47.5	35.7	40.8	50.4	16.2	24.5	29.3	**33.3**	31.2	66.9
	WE-LSTM	**72.7**	**69.5**	**71.1**	**56.2**	**48.3**	**52.0**	**60.4**	**31.1**	**41.1**	**61.1**	29.4	**39.7**	**53.6**

Table 3.4: Punctuation restoration results for Hungarian ASR transcripts

ASR Transcript	**Model**	**Comma**			**Period**			**Question**			**Exclamation**			**SER**
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(19,19)	**64.5**	55.8	59.9	41.1	31.2	35.6	**41.2**	8.8	14.4	48.8	17.1	25.4	79.2
	WE-BiLSTM	63.9	**67.7**	**65.7**	**50.5**	**49.0**	**49.8**	37.7	**24.1**	**29.4**	**60.9**	**24.0**	**34.4**	**70.1**
On-line mode	MaxEnt-(25,1)	**64.3**	54.9	59.2	38.9	29.4	33.5	36.0	7.1	11.9	47.1	20.6	28.6	81.3
	WE-LSTM	63.8	**65.1**	**64.4**	**47.8**	**42.0**	**44.7**	**48.5**	**20.5**	**28.9**	**61.8**	**21.7**	**32.1**	**73.1**

As it can be seen, the prediction results for comma stand out from the others for all methods and configurations. This can be explained by the fact that Hungarian has generally clear rules for comma usage. Nevertheless, the prediction of sentence boundary markers (especially period) may also benefit from acoustic-prosodic information. This assumption will be supported by the results presented in Section 3.5.4.

As Table 3.3 shows, switching to the RNN-based punctuation restoration for Hungarian

manual transcripts significantly ($p < 0.05$) reduces SER by around 20% relative compared to the baseline MaxEnt approach. WE-BiLSTM and WE-LSTM are especially powerful in restoring periods, question marks and exclamation marks as they are able to exploit large contexts much more efficiently than the MaxEnt taggers. These results on manual transcripts serve for benchmarking the punctuation restoration performance on ASR transcripts, which corresponds to the real-life use case, i.e. automatic closed captioning.

As outlined in the Section 3.1.1., limiting the future context and propagation of ASR errors into the punctuation recovery pipeline are considered to be the most important factors hindering effective recovery of punctuations in live TV streams. Comparing off-line and on-line results, shows however, that a large future context is less crucial for robust punctuation restoration, contradictory to my expectations: dropping the future context causes only 2-4% relative decrease in performance. The features from the future word sequence seem to be more related to increase recall, otherwise the WE-LSTM is an equally suitable model for punctuation recovery, and seems suitable for real-time application.

Comparing results on manual and ASR transcripts shows that ASR-errors propagate into the punctuation module: switching from manual transcripts to ASR hypotheses resulted in 15-20% increase in SER (see Table 3.4). WER is reflected in the increased SER, as word errors may destroy punctuation slots. Although the performance gap is decreased between the two approaches when evaluating ASR transcripts as test input, my RNN approaches still outperform the MaxEnt baseline significantly ($p < 0.05$), by 10% relative w.r.t SER.

3.2.4 Genre Analysis for Word-based Punctuation

The test part of Hungarian Broadcast Database can be divided into 6 subsets based on the genres of the transcripts (see Table 3.2). I also analysed punctuation recovery for these subsets, hypothesizing that more informal and more spontaneous genres are harder to punctuate, in parallel to the more ASR errors seen in these genres. Some of the punctuation marks for specific genres were dropped ("N/A" in Table 3.2), when the punctuation mark is not used in such genres (e.g. Weather Forecast transcripts do not contain any question mark).

As the RNN models outperformed the MaxEnt taggers for every genre, only the results of WE-BiLSTM and WE-LSTM systems are shown in Table 3.5 and 3.6.

Comparing the results to the statistics in Table 3.2, it is visible that the number of training samples has positive influence on the performance of the punctuation recovery system: performance is higher for genres having more training samples (BC-news, BC-conversations,

Table 3.5: Hungarian manual transcript results by genres

Manual Transcript	Genre	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
RNN Off-line mode	Weather	61.2	54.3	57.5	46.7	46.7	46.7	N/A	N/A	N/A	**90.0**	45.0	60.0	69.3
	BC-News	**89.9**	**84.4**	**87.1**	**84.3**	**90.7**	**87.3**	**91.7**	**50.0**	**64.7**	83.9	**56.5**	**67.5**	**20.0**
	Sport news	68.3	60.6	64.2	49.4	51.4	50.4	N/A	N/A	N/A	75.0	30.0	42.9	67.0
	BC-Conv.	80.4	74.5	77.3	63.9	64.9	64.4	63.0	46.4	53.5	88.9	18.5	30.6	38.7
	Sport mag.	61.2	61.1	61.1	43.9	49.3	46.5	55.2	37.5	44.7	38.5	9.4	15.2	73.1
	Magazine	67.6	67.6	67.6	45.1	46.3	45.7	50.5	29.7	37.5	50.0	5.6	10.1	58.6
RNN On-line mode	Weather	60.2	57.5	58.8	45.7	37.9	41.4	N/A	N/A	N/A	87.5	35.0	50.0	70.6
	BC-News	**88.4**	**83.1**	**85.7**	**86.6**	**81.3**	**83.9**	**75.0**	40.9	**52.9**	**100.0**	**67.4**	**80.5**	**24.1**
	Sport news	68.7	57.2	62.4	42.4	37.5	39.8	N/A	N/A	N/A	90.0	60.0	72.0	74.2
	BC-Conv.	80.1	74.0	76.9	66.7	54.8	60.1	63.0	**45.6**	**52.9**	77.6	29.2	42.5	40.8
	Sport mag.	60.8	59.7	60.3	42.3	34.8	38.2	53.3	38.3	44.5	20.0	7.5	11.0	77.3
	Magazine	67.6	65.1	66.3	43.5	32.8	37.4	57.3	27.2	36.9	36.4	11.3	17.2	61.5

Table 3.6: Hungarian ASR transcript results by genres

ASR Trans.	Genre	Comma			Period			Question			Exclamation			SER	WER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1		
RNN Off-line mode	weather	57.8	54.3	55.9	48.1	43.8	45.8	N/A	N/A	N/A	**100.0**	**60.0**	**75.0**	74.3	6.8
	BN	65.7	**82.2**	**73.0**	**63.5**	**82.7**	**71.8**	**52.9**	**40.9**	**46.2**	75.0	45.7	56.8	**57.4**	10.1
	sport news	53.3	55.9	54.6	37.9	41.1	39.5	N/A	N/A	/NA	**88.2**	**53.6**	**66.7**	93.1	21.4
	BC	**70.4**	68.8	69.6	55.6	49.4	52.3	44.1	28.4	34.5	70.8	13.1	22.1	59.6	24.7
	sport magazine	52.8	58.0	55.3	37.5	41.1	39.2	42.9	18.8	26.1	N/A	N/A	N/A	93.2	30.3
	magazine	59.6	59.9	59.7	34.6	29.1	31.6	31.9	13.9	19.4	16.7	1.4	2.6	82.3	38.7
RNN On-line mode	weather	61.4	57.9	59.6	43.1	39.1	41.0	N/A	N/A	N/A	**88.9**	40.0	55.2	73.2	6.8
	BN	64.8	**80.6**	**71.8**	**62.4**	**73.5**	**67.5**	40.0	9.1	14.8	75.0	45.7	56.8	**61.8**	10.1
	sport news	52.8	54.0	53.4	35.4	34.6	35.0	N/A	N/A	N/A	83.3	**53.6**	**65.2**	96.7	21.4
	BC	**70.2**	67.1	68.6	53.5	40.9	46.3	**46.4**	**25.6**	**33.0**	69.2	13.8	23.1	62.4	24.7
	sport magazine	53.1	55.4	54.2	35.6	30.9	33.1	41.4	22.7	29.3	16.7	5.7	8.5	94.0	30.3
	magazine	58.5	59.9	59.2	36.0	22.3	27.6	46.3	12.0	19.1	42.9	4.2	7.7	83.2	38.7

magazines). The relatively large difference in SER among these three, well-modelled genres point out another important factor, which is the task predictability. Analogous to language modelling, the more formal the task is, the better is the predictability of punctuation (see BC-news results). Indeed, conversational (BC-conversations) and informal (magazine) speech styles (which tend to neglect the usual word-order constraints, hence using more ungrammatical phrases and increased number of disfluency events) make prediction more difficult and introduce more punctuation errors compared to more formal styles.

The relatively high SER of the weather forecast and the sport program genres points out the importance of using sufficient amount of in-domain training data.

By comparing SER of the manual and ASR transcripts, some interesting conclusions can be drawn. For the well-modelled genres (BC-News, BC-Conv., magazine) the positive linear correlation of SER and WER is noticeable. However, for the remaining genres (weather, sport news, sport magazine), this relationship between SER and WER is much less predictable. The connection between these objective metrics are further discussed in Section 4.2.3.4. It is

particularly difficult to explain the relatively poor results for the sport news genre. Whereas the WER of the ASR transcript is moderate for Hungarian (21.4%), there is an almost 40% increase in the SER of punctuation. I assume that this phenomenon is related to the high number of named entities in the sport news program, considering that the highest OOV Rate (10%) can be spotted for this genre among all the 6 tested genres (OOVs result in the *Unknown* word vector in the punctuation model). Moreover, the high ratio of unforeseen expression patterns compared to the content of training data also can have negative impact on the test results.

After the word-level RNN model, I also propose a character-level CNN-RNN punctuation approach in the next section. The advantage of the character-level model is that it is not necessary to map OOV words to a common *Unknown* symbol, as the vocabulary will not explode thanks to the limited number of characters, and the influence of rare words are not lost. For instance, if there is a specialized text (e.g. medical report, audiobook), there may be a topic mismatch with the (pre-trained) word embedding; a character-level model is more likely to deal with such situations as well.

The below example compares a Hungarian "gold" transcript with an automatically punctuated ASR transcript. This short paragraph (about a religious mission) reflects on the better predictability of comma in some cases (before *hogy* and *mert* conjunction words), and the typical error situations of period (substitution error with comma, or insertion error by closing the sentence at a meaningful part). I marked the deletion error with "*" and the missing, correct punctuation mark.

REFERENCE TRANSCRIPT:
Lovas zarándoklat indul holnap a svábhegyi Anna-rétről a csíksomlyói búcsúba.
A lovasok a nyolcszáz kilométert két csoportban teszik meg, hogy két zarándokútvonalat is bejárhassanak.
A túra Erdélyben ér össze, együtt érkeznek Csíksomlyóra.
A túra célja, hogy pénzt gyűjtsenek a Normafa mögötti Anna-réten építendő kápolnára.
Csabi és Niké, a lovas zarándoklat főszereplői.
Már több ezer kilométer van mögöttük, hiszen már négyszer megjárták a csíksomlyói búcsút.
A két ló és lovasaik a Világ Győzedelmes Királynője nevet viselő zarándokúton indulnak el csütörtökön.
Hálózsák, nyeregtáska, ebbe jön két katonai poncsó.
Összesen ennyit vihetnek a résztvevők a túrára.
Heti hat napot mennek, aztán egyet pihennek, mert az idősebb lovaknak kell a szabadnap.

AUTOMATIC TRANSCRIPT:
Lovas zarándoklat indul holnap a svábhegyi Anna rétről a csíksomlyói búcsúba.
A lovasok a nyolcszáz kilométert két csoportban teszik meg, hogy két zarándok-útvonalat is bejárhassanak.
A túra Erdélyben ér össze, együtt érkeznek Csíksomlyóra.
A túra célja, hogy pénzt gyűjtsenek a Normafa mögötti Anna réten.[Insertion]
építendő kápolnára*.[Deletion]
Csabi és még ki*,[Deletion] a lovas zarándoklat főszereplői.
Már több ezer kilométer van mögöttük, hiszen már négyszer megjárták a csíksomlyói búcsút.
A két ló és lovasaik a Világ Győzedelmes Királynője nevet viselő zarándokúton indulnak el.[Insertion]
csütörtökön*.[Deletion]
hálózsák*,[Deletion] nyeregtáska, ebbe jön két katonai poncsó, [Substitution]
összesen ennyit vihetnek a résztvevők a túrára.
Szóló heti hat napot mennek, aztán egyet pihennek, mert az idősebb lovaknak kell.[Insertion]
a szabad nap*.[Deletion]

3.3 The Character-based CNN-RNN Model

In case of my character-level neural network, also fixed-length chunks are used. As Fig. 3.6 shows, each word is represented as a sequence of characters.

Each character-sequence is padded to obtain a fixed length input, using a *Padding* token, then each character is mapped into the embedding space. Pre-trained embeddings are not available for characters, unlike for word-level models. After the character-level embedding transformation, first a 1D-convolution operation (using numerous convolution filters with different weights, "striding" through the sequences) produces many feature maps. Then a downsampling with MaxPooling is executed to extract a new feature vector. Finally, a BiLSTM is responsible again for representing the context of x_t. At the output, a softmax layer determines the posterior probabilities over the output classes for the punctuation task. The intermediate Dropout layers help preventing overfitting on the training set [150]. This model is called CE-CNN-BiLSTM, where "CE" comes from Character Embedding.

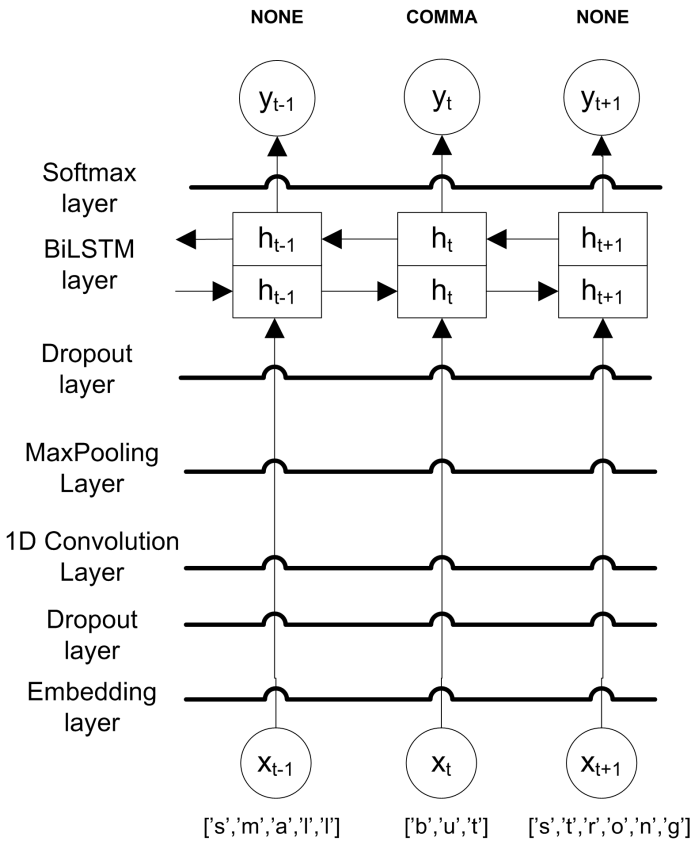

Figure 3.6: Structure of CE-CNN-BiLSTM RNN model

3.3.1 Hyperparameters

I performed a systematic grid search optimization for hyperparameters on the same validation set for CE-CNN-BiLSTM, as I did it earlier for WE-BiLSTM. The length of chunks, vocabulary size, number of hidden states, mini-batch size, optimizers were investigated for both models, while additionally, the length and the number of convolution filters, the stride, and the size of the MaxPooling window were fine-tuned for character-level models. I also used early stopping on the validation set to prevent overfitting, controlled with a *patience* variable. Table 3.7 summarizes the final values of each hyperparameter used in Hungarian CE-CNN-BiLSTM model.

Table 3.7: Hyperparameters of WE-BiLSTM and CE-CNN-BiLSTM models for Hungarian

Language	Model	Chunk Size	Vocab. Size	Embedding dimension	#LSTM cells	Batch size	Optimizer	Filter length	#Filters	Stride	Size of MaxPooling Window	Patience
HUN	WE-BiLSTM	200	100,000	152|300|600	**512**	**128**	**RMSProp**	N/A	N/A	N/A	N/A	3
	CE-CNN-BiLSTM		100	80				6	70	2	25	

Note: As Table 3.7 shows, I kept most of the values of WE-BiLSTM (seen in Table 3.1), except investigating other pre-trained embeddings provided by the author of [142], to decrease the size of the model without significant performance loss if it is possible. I evaluated my approach with an 152-dimensional GloVe and a 300-dimensional Word2vec word embedding. In the notation of word-level Hungarian models, the number refers to the embedding dimension (e.g. WE-BiLSTM-152). The previously presented word-level model (with 600-dimensional embedding) was also re-trained from scratch.

3.3.2 Experimental Results for Character-based Punctuation

Thesis II.B. [C9] *I have experimentally confirmed, that automatic punctuation using a character-based CNN-RNN model for Hungarian is also possible; my CNN-RNN model yields slightly lower performance compared to my word-based punctuation approach presented in Thesis II.A.*

Despite its lower performance, the character-level model has some advantages which are worth to consider: it eliminates the issue of vocabulary explosion, and therefore it is relevant for the highly agglutinating Hungarian language. In addition, it has considerably smaller footprint than the word-level model. If the storage consumption is a key factor to integrate a punctuation module to an ASR-based application, halving the dimension of word embedding from 600 to 300 yields a cca. half-sized model (250 MB vs. 130 MB), but switching from

word-level to character-level results in even more saving by a 14:1 ratio (130 MB vs. 9 MB)! The price of this profit is reflected in a longer training period as the character-level model needed 3-times longer training than the word-level model to reach the optimum in the validation loss (65 epochs during 9 hours vs. 13 epochs during 3 hours). Models were trained on NVIDIA Titan X GPU, assisted with a quad-core Intel Core i5-6600K CPU @ 3.5 GHz and 16 GB system memory.

The obtained results are presented in Table 3.8 for the manual transcripts and in Table 3.9 for the ASR transcripts, respectively. Considering the performance of character-based punctuation recovery by the different punctuation marks, the tendency is the same as it was shown in Section 3.2. Therefore, only the difference between the word-level and character-level methods is emphasized. The importance of word-level features is reflected by the difference in character- and word-level results; especially in case of period and question mark restoration. For instance, as the "Wh-" words mostly occur in the context of a question in English (also supported with specific word order), the Hungarian language has its well-identifiable interrogative terms, such as *Miért? <-> Why?* or *Hol? <-> Where?*.

Table 3.8: Punctuation restoration results for Hungarian manual transcripts

Model Type	Model name	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Char-level	CE-CNN-BiLSTM	74.1	65.9	69.8	56.2	44.9	49.9	54.5	28.6	37.5	62.3	20.0	30.3	55.0
Word-level	WE-BiLSTM-152	73.7	67.5	70.5	61.8	46.8	53.3	62.0	24.6	35.2	46.3	23.0	30.7	52.0
	WE-BiLSTM-300	74.5	69.8	72.1	60.4	54.2	57.1	47.5	**43.8**	45.6	45.3	24.4	31.7	49.7
	WE-BiLSTM-600	74.5	70.5	72.4	61.8	52.5	56.8	56.8	32.2	41.1	50.8	**31.7**	**39.0**	49.2

Table 3.9: Punctuation restoration results for Hungarian ASR transcripts

Model Type	Model name	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Char-level	CE-CNN-BiLSTM	64.9	61.8	63.3	46.8	38.9	42.5	40.2	18.0	24.9	65.9	16.0	25.7	73.7
Word-level	WE-BiLSTM-152	65.6	63.5	64.5	54.4	40.1	46.2	47.5	13.4	20.9	72.3	17.1	27.7	68.8
	WE-BiLSTM-300	65.8	65.5	65.6	52.0	46.4	49.1	37.9	**27.7**	**32.0**	67.9	15.1	24.8	68.3
	WE-BiLSTM-600	64.7	67.3	66.0	53.5	44.9	48.8	45.9	18.0	25.9	74.7	**21.1**	**33.0**	68.3

Turning to results on ASR transcripts in Table 3.9, ASR errors propagate and cause more punctuation errors, finally translated in an overall 20% performance decrease in SER. In details, the comma restoration remained the most precise, followed by period. Furthermore, the recovery accuracy of question and exclamation mark are almost identical.

The performance of the character-level model is slightly worse compared to the word-level one (with different pre-trained embeddings), by 5-10% relative w.r.t. SER in the case of manual transcripts, and by 7% relative w.r.t. SER in the case of ASR transcripts. When the pre-trained embeddings are not used for the word-level model, the training corpus is the same

for the word-level and character embeddings. Setting the *embedding_size* hyperparameter to 600 (dimension), I measured $SER = 51.3\%$ in the case of manual transcripts, and $SER = 69.5\%$ in the case of ASR transcripts. The performance gap decreases between the character-level and word-level models (rel. 7% and rel. 6%), but still the word-level model shows better result in the punctuation restoration.

The reason for creating a separate thesis in II.B from the character-level punctuation model will hopefully become obvious in Section 3.4. My next goal was to exploit the character-level model (by the handling of rare words) in a combination with word-level models as a textual hybrid and an ensemble approach, assuming higher performance in punctuation accuracy. These combined methods and their experimental results are described in the following sections.

3.4 Combined Text-based Approaches

3.4.1 Textual Hybrid Model

The text-based hybrid model receives multiple inputs; it learns both from character- and word-level features, and produces output probabilities for punctuation marks. The hyperparameters of the best character- and word-level models were re-used. As Fig. 3.7 shows, the lower stacked layers of the character- and word-level models are concatenated by the Merge layer. Therefore, a joint feature vector is created from character-level features (learned from 1D Convolution - MaxPooling operations) and pre-trained word-level embeddings (note: the size of feature vectors is equal to the chunk length). Then the BiLSTM and softmax layers play the same role as in the case of single models; finally, the hybrid model makes predictions for punctuation marks.

3.4.2 Textual Weighted Ensemble Model

The ensemble consists of two single models; a character-level and a word-level one. The two models are trained separately, and only their output is combined (unlike the hybrid model, where the inputs are combined into one model). After the single models have already been trained on the available data, the word-level and character-level predictions are combined by taking the weighted average of their output probabilities over punctuation marks, where the weights sum to 1. The weights were determined on the validation set, then the same weights were applied to the manual and ASR transcripts of the test set. In the notation of a weighted ensemble, the first weight belongs to the character-level model, and the second

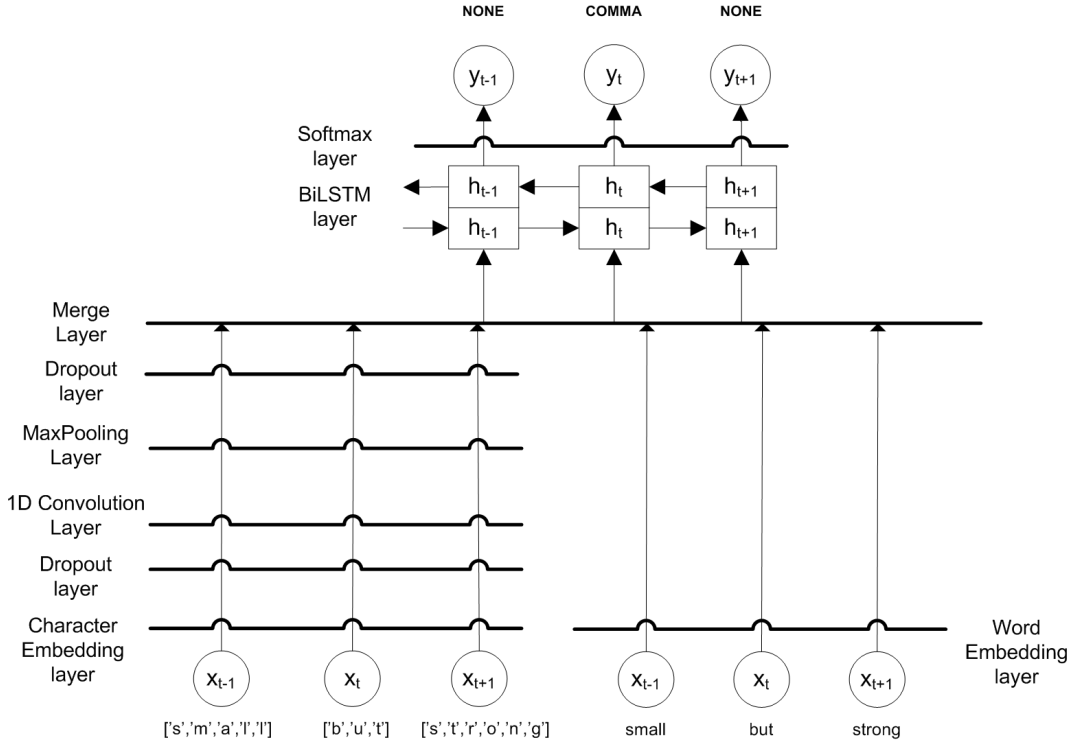

Figure 3.7: Structure of the textual hybrid model

one to the word-level model.

3.4.3 Experimental Results for Combined Text-based Punctuation

Thesis II.C. [C9] *I have experimentally confirmed, that my hybrid punctuation model consisting of character- and word-based components significantly outperforms my word-based punctuation model for Hungarian in terms of SER (by 3-9% relative for manual transcripts and by 2-3% relative for ASR transcripts on the investigated corpus).*

The obtained results are presented in Table 3.10 for the manual transcripts and in Table 3.11 for the automatic (ASR) transcripts, respectively. For the sake of completeness and better understanding of the notations, the results of the base models (both character-level and word-level) are also included in the tables.

Comparing the hybrid and weighted ensemble models to the character- and word-level approaches, the combined models yield better performance than their base models with a "single" (word or character) input. The weighted ensembles provide the highest precision in most cases and the weights show the moderate dominance of the word-level model in the

Table 3.10: Punctuation restoration results for Hungarian manual transcripts

Model Type	Model name	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Char-level	(1) CE-CNN-BiLSTM	74.1	65.9	69.8	56.2	44.9	49.9	54.5	28.6	37.5	62.3	20.0	30.3	55.0
Word-level	(2) WE-BiLSTM-152	73.7	67.5	70.5	61.8	46.8	53.3	62.0	24.6	35.2	46.3	23.0	30.7	52.0
	(3) WE-BiLSTM-300	74.5	69.8	72.1	60.4	54.2	57.1	47.5	**43.8**	45.6	45.3	24.4	31.7	49.7
	(4) WE-BiLSTM-600	74.5	70.5	72.4	61.8	52.5	56.8	56.8	32.2	41.1	50.8	**31.7**	**39.0**	49.2
Hybrid	(1)+(2)	76.1	70.3	73.1	60.7	**59.6**	**60.1**	63.0	40.9	**49.6**	43.7	25.2	32.0	47.3
	(1)+(3)	74.6	72.6	**73.6**	**65.5**	53.5	58.9	61.9	37.1	46.4	45.6	26.2	33.3	**46.7**
	(1)+(4)	74.3	**72.7**	73.5	63.6	53.4	58.1	63.9	30.8	41.6	48.3	30.6	37.5	47.5
Weighted Ensemble	0.47*(1)+0.53*(2)	76.0	66.9	71.2	63.6	47.0	54.1	**69.6**	27.0	38.9	**67.3**	20.4	31.3	50.4
	0.35*(1)+0.65*(3)	**76.5**	69.3	72.7	63.5	52.7	57.6	56.1	38.9	45.9	46.4	22.4	30.2	47.9
	0.38*(1)+0.62*(4)	76.3	69.5	72.7	64.3	51.3	57.1	62.7	31.9	42.3	55.8	26.2	35.7	48.0

Table 3.11: Punctuation restoration results for Hungarian ASR transcripts

Model Type	Model name	Comma			Period			Question			Exclamation			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Char-level	(1) CE-CNN-BiLSTM	64.9	61.8	63.3	46.8	38.9	42.5	40.2	18.0	24.9	65.9	16.0	25.7	73.7
Word-level	(2) WE-BiLSTM-152	65.6	63.5	64.5	54.4	40.1	46.2	47.5	13.4	20.9	72.3	17.1	27.7	68.8
	(3) WE-BiLSTM-300	65.8	65.5	65.6	52.0	46.4	49.1	37.9	**27.7**	**32.0**	67.9	15.1	24.8	68.3
	(4) WE-BiLSTM-600	64.7	67.3	66.0	53.5	44.9	48.8	45.9	18.0	25.9	74.7	**21.1**	**33.0**	68.3
Hybrid	(1)+(2)	65.9	66.1	66.0	51.7	**50.8**	**51.2**	46.2	22.0	29.8	63.2	13.7	22.5	67.6
	(1)+(3)	65.0	68.3	66.6	**56.3**	45.4	50.2	48.3	23.0	31.2	67.4	16.6	26.6	**66.2**
	(1)+(4)	65.0	**68.9**	**66.9**	54.6	46.0	49.9	**50.5**	18.4	27.0	69.2	**21.1**	32.4	67.0
Weighted Ensemble	0.47*(1)+0.53*(2)	**67.6**	62.7	65.0	55.4	39.0	45.8	49.4	15.2	23.2	**76.0**	16.3	26.8	67.5
	0.35*(1)+0.65*(3)	67.3	62.9	65.0	53.4	41.6	46.6	45.6	19.5	27.3	73.0	16.6	27.0	68.0
	0.38*(1)+0.62*(4)	66.9	63.6	65.2	54.1	41.4	46.9	47.8	17.7	25.8	74.7	17.7	28.6	68.0

ensemble. On the other hand, the hybrid models (involving the character-level features for the punctuation task) yield higher recall, and the best overall results, accounting for 3-9% significant ($p < 0.05$) improvement in SER compared to their base word-level models with various word embeddings.

Comparing the hybrid and the weighted ensemble to the word-level approach, the combined models give a somewhat smaller, but still significant ($p < 0.05$) 2-3% relative in performance w.r.t SER when using ASR transcripts. All in all, the tendency does not change, thus the hybrid model (trained with character- and word-level features) shows the best results in punctuation restoration.

As acoustic features are also known to be exploitable for punctuation, the next step is to investigate the combination of the textual and acoustic features to predict punctuation marks.

3.5 Towards a Hybrid Textual-acoustic Approach

3.5.1 Speech Materials

Unfortunately, the Hungarian Broadcast Dataset was not applicable for the textual-acoustic experiments in its original form; the main reason was that the manual transcripts were

imprecise in some parts compared to the audio recordings (and rather contained stylistic changes), which resulted in incorrect or unsuccessful extraction and alignment of the acoustic features.

Instead of that, first I used two databases for the evaluation of Hungarian hybrid textual-acoustic RNN approaches with precise transcriptions. In the first dataset, the audio was associated with a dataset covering mainly Hungarian Broadcast News genres [151] of 3.5 hour recordings with 23K words. Secondly, the Hungarian BABEL [59], a read speech database recorded from non-professional native speakers was used for the experiments, containing cca. 3 hours of recordings and 22K words. The corresponding results can be found in [C13, C18]. The main problem with these datasets, that due to the highly error-prone circumstances (WER=34,6% and WER=50% for the two corpora), the role of punctuation may become irrelevant.

Finally, I organized a sub-corpus from the original test dataset of the Hungarian Broadcast Dataset, involving only the reference and ASR transcripts of all weather forecast, all sport news and some broadcast news (selected based on the lowest WER values). This corpus contains cca. 3 hours of recordings and 24K words, further referred to as **MTVA-3h**. I listened to all of these recordings and corrected the manual transcripts so that it truly reflects the underlying word sequence. The ASR system produced a WER=10.3% on this part of the Hungarian Broadcast Dataset. Unlike in the case of the original dataset, I carefully annotated the ASR transcripts of this small corpus with punctuation marks. Again, the number of question and exclamation marks are underrepresented in this sub-corpus. This material was divided as 60%-20%-20% for training, validation and testing purposes.

3.5.2 Acoustic-prosodic Model

Finally, I was interested in exploiting the advantages of text- and acoustic-based punctuation methods together, which required to integrate an acoustic-prosodic component. In [83], the authors evaluated punctuation based on phonological phrase (PP) segmentation in Hungarian. That approach involves a separate HMM to obtain the phonological phrases based on prosodic features (it is the same method which was presented in Section 2.3).

Unfortunately the combination with my text-based punctuation approaches did not lead to satisfactory results, most likely caused by the mismatch between the processing units: in most cases phonological phrases contain several words, not just a single one. In my previous punctuation models, the word was the base component, i.e. the word embeddings and the character embeddings as fixed-length sequences derived from words.

To keep aligned with the word-level concept, I tried to incorporate the posterior values (soft labels) of the acoustic-prosodic RNN punctuation model by [83]. One of the main drawbacks of this concept is that information about intermediate words (where the phonological phrase boundaries are not detected) is lost, i.e. missed slots are generated, where punctuation marks are omitted. The second one is that there could be false PP boundaries as well. As a consequence, the missing posterior information and mismatch between the phrase-word alignment had distorted the knowledge available at the character-level and word-level, and it caused deterioration in results. When I used features as the combination of word position within the phrase and various PP types (e.g. begin-descending) instead of the posterior information, the results were still below the expectations. These solutions showed that the two (textual and prosodic) systems could not function in complementing each other. Although I didn't involve my PP detector to my punctuation experiments, the same problems could come up in the case of the WCAD approach.

To overcome these difficulties and also to ensure an end-to-end solution where most of the modelling job is done by a neural network, I preserve only the acoustic-prosodic feature pre-processing steps from [83]. F0 and overall energy was extracted (for energy I use a 150 ms long window and do not decompose it into mel bins) by 10 ms frame rate and I smoothed the signals by applying a 5-points median filter. First and second order derivatives were also computed, obtained by approximating the derivatives d_t of x_t in a $W = 30$ frame context by the following regression formula:

$$d_t = \frac{\sum_{i=1}^{W/2} i(x_{t+i} - x_{t-i})}{2\sum_{i=1}^{W/2} i^2} \tag{3.2}$$

Wherever a word boundary is hypothesized by the ASR (or the alignment of the transcripts), two 15 frame long portions of this 6-dimensional feature sequence are extracted preceding and following the word boundary. Statistics composed of the minima, maxima and mean of the 6 values are computed for the two portions separately and added to the input vector of the punctuation module. The input vector is augmented by the durations of the preceding word and of the pause at the word boundary. This input vector is fed into a bidirectional LSTM layer, followed by a softmax to predict the punctuation. This P-BiLSTM model is kept small (see Table 3.12) to allow for use with low amount of training data as I have a limited amount of audio data transcribed with punctuation.

3.5.3 Textual-acoustic Hybrid Models

The acoustic-based model is trained "from scratch" on **MTVA-3h** corpus, while the character- and word-level models are pre-trained with the corpus described in Section 3.2.1. Thereafter, further training is carried out with transfer learning method [152] on the smaller **MTVA-3h** corpus. Then I combined the character (C, earlier CE-CNN-BiLSTM), word (W, earlier WE-BiLSTM) and prosody (P from P-BiLSTM) based models pairwise (C+W, C+P, W+P) and also into a triple (C+W+P) hybrid model.

In case of these hybrid models, the most efficient combination was to concatenate the respective BiLSTM hidden states of C and/or W models with the input of the P model. Compared to the text-based solution presented in Section 3.4, these hidden states were not utilized in that final model. Finally, a new shared bidirectional LSTM layer and a new softmax layer are stacked over the concatenated layers of the single models at the textual-acoustic approaches. The hyperparameters of the individual models are summarized in Table 3.12.

Table 3.12: Hyperparameters of the individual models for Hungarian

Input	Model	Chunk Size	Vocab. Size	Embedding dimension	#LSTM cells	Batch size	Optimizer	Filter length	#Filters	Stride	Size of MaxPooling Window	Patience
Words	W	200	100,000	300	512	128	RMSProp	N/A	N/A	N/A	N/A	3
Chars	C	200	100	80	512	128	RMSProp	6	70	2	25	3
Prosody	P	200	N/A	N/A	512	16	RMSProp	N/A	N/A	N/A	N/A	3

3.5.4 Experimental Results for Punctuation with Textual-acoustic Hybrid Models

Thesis II.D. [C13, C18] *I have experimentally confirmed, that my hybrid punctuation model consisting of character-, word- and prosody-based components significantly outperforms my character-word based hybrid on the manual and ASR transcripts of the involved Hungarian corpus (by 8% and 6% relative in SER, respectively).*

I used again the F1-score computed from precision and recall for evaluation (scaled in $0-100$ range), altogether with Slot Error Rates (SER). The results for Hungarian with **MTVA-3h** corpus are shown in Fig. 3.8 and in Table 3.13. Text-based (C and W) models yield good comma recovery, but the C model performs weaker for periods, both on reference (REF) and ASR transcripts. On the other hand, as sentence endings tend to be more marked by prosody, the P model performs well with period prediction, but has weak comma prediction

capabilities. The W model is only slightly worse in the period restoration than P model is. The reason behind that is supposed to be two-fold; (i) Results reflect to my previous genre-level analysis, where I showed, that the performance of punctuation restoration depends on the task predictability. According to the distribution of genres in this sub-corpus, nearly 60% of the samples belong to the broadcast news, which is the most formal category, and it compensates the weaker performance for weather forecast and sport news as well.

Combining the single models leads to overall improvement (with C+P, W+P and C+W+P). Overall the best models were C+W+P on reference ($F1 = 78.0\%; SER = 35.0\%$) and also C+W+P on ASR ($F1 = 72.8\%; SER = 44.1\%$). Involving prosodic features (W+P), I obtained significant ($p < 0,05$) improvement over the word-level baseline by 6% and 7.8% on the reference and the ASR transcripts, respectively. Similarly, when I added prosodic features to the C+W hybrid (hence C+W+P was established), I measured significant ($p < 0,05$) improvement in punctuation performance over the C+W hybrid by 8.4% and 5.6% on the reference and the ASR transcripts, respectively.

Table 3.13: Slot Error Rates for the MTVA-3h Corpus

Model Type	**SER REF**	**SER ASR**
C	51.0	57.4
W	38.5	48.7
P	76.7	84.6
C+W	38.2	46.7
C+P	36.3	45.2
W+P	36.2	44.9
C+W+P	35.0	44.1

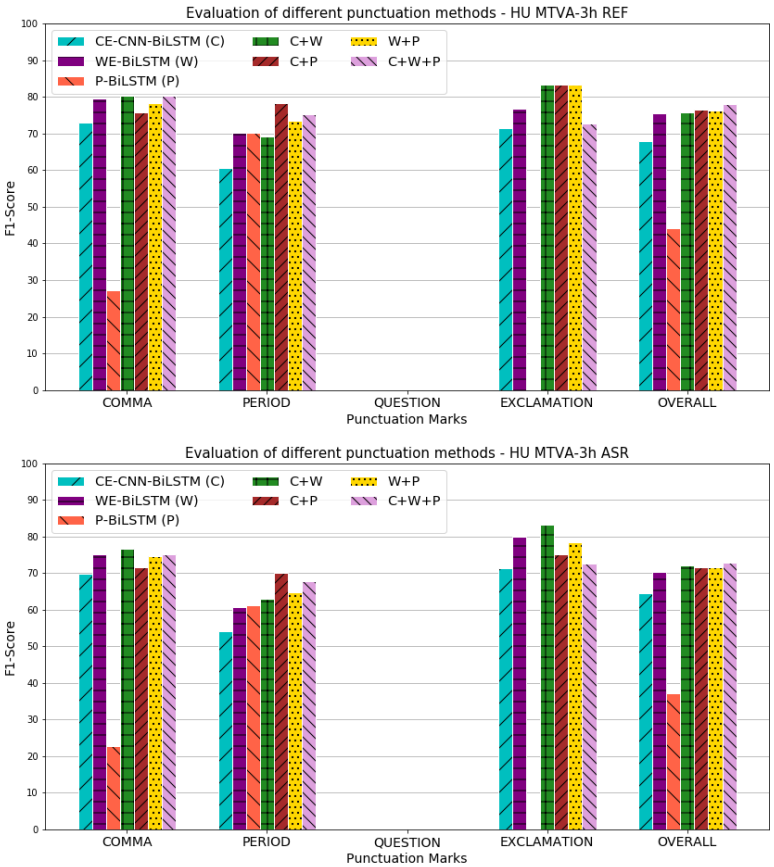

Figure 3.8: F1-results on reference and ASR transcripts of MTVA-3h corpus, including hybrid models for punctuation

Regarding the restoration of question and exclamation marks, the P model performs weak on **MTVA-3h**. The reason behind that the intonation of the questions and exclamations often does not correspond to the interrogative or exclamative modalities, but they are realized with declarative patterns. This was confirmed by listening the particular utterances, but it has been also observed in [83] and [110] – despite prosody should theoretically be an ideal candidate to detect questions, practice does not confirm this. The confusion matrices showed also the substitution errors between period and question or exclamation marks. Although, the data imbalance is high (the proportion of question and exclamation marks was only 0.6% and 2.6% on **MTVA-3h**), the textual models can deal with exclamation marks, thanks to the pre-trained phase as well.

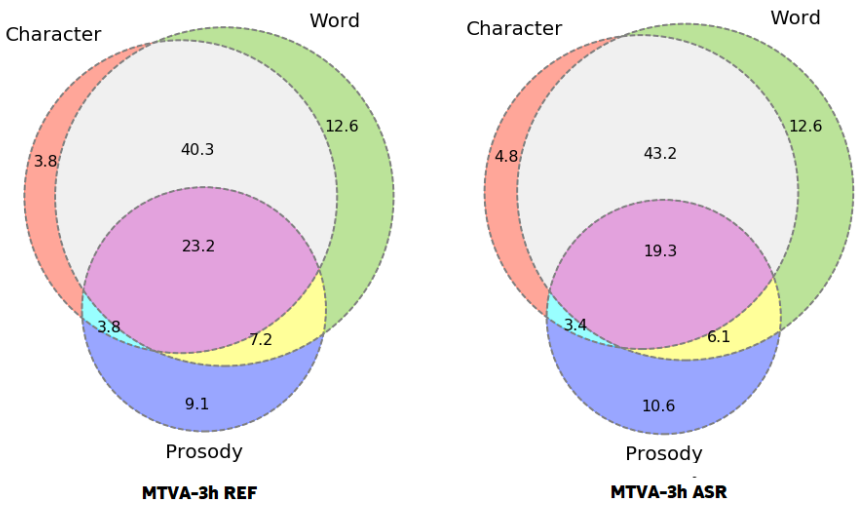

Figure 3.9: Coverage of the Hungarian individual punctuation models

In Fig. 3.9, Venn-diagrams show the share (in %) of the individual models on correctly recovered punctuation marks in Hungarian, i.e. 100% equals to all correctly recovered punctuation marks by any of the models. The dominance of the text-based models can be observed on the reference transcripts. Although the P model recovers the least punctuation marks overall on **MTVA-3h**, it has still an important role. It is able to capture cca. 9% of punctuations on reference transcripts, which are not seen by any other model. When working on ASR output, as expected, the amount of slots detected by C and P slightly increases, showing that C and P models are more robust on ASR transcripts. The relative differences of the single models in SER between reference and ASR transcripts also refers to that. Nevertheless, word errors have a negative impact on P model, if word boundaries are falsely hypothesised.

Summarizing the results, for the highly agglutinating and relatively free word order Hungarian, significant improvement (by $p < 0.05$) can be obtained in overall punctuation over the W baseline by adding the character-level and prosodic features.

3.5.5 English Experiments for Automatic Punctuation

3.5.5.1 Speech Material

During my automatic punctuation studies, I evaluated my models for English as well, described in the following publications: [C6, C7, C9, C13, C17]. I used an English dataset which consists of IWSLT TED Talk transcripts, which is a commonly used benchmark dataset for English punctuation models [75, 76, 91, 95]. I used the predefined train, validation and test sets, containing 2.1M, 296k and 13k words respectively, and dealing with three types of punctuations (comma, period and question mark). This corpus contains no audio, hence audio was derived from the IWSLT2011 talk translation dataset [153] containing 6 hours of speech. This database was used for training and testing textual-acoustic hybrids.

3.5.5.2 Word-level Experiments

Firstly, I compared my word-based RNN models to the state-of-the-art punctuation recovery systems [75, 76], in both on-line and off-line modes [C6, C7, C17]. The structure of these models was depicted in Section 3.2; they contain the same layers like the Hungarian models. To recall, basically the on-line and off-line modes depend on the type of the LSTM hidden cells, i.e. they are uni- or bidirectional, respectively. A 100-dimensional pre-trained GloVe word embedding was also used in order to represent the English words. The hyperparameters of the English models are shown in Table 3.14.

Table 3.14: Hyperparameters of WE-BiLSTM and WE-LSTM models for English

Language	**Model**	**Chunk Length (#words)**	**Vocab. Size (#words)**	**Word Embedding dimension**	**#Hidden states**	**Batch size**	**Optimizer**	**Patience**
EN	WE-BiLSTM	200	27,244 (by [75])	100 (by [75])	256	128	RMSProp	2
EN	WE-LSTM	250						

As it is shown in Tables 3.15 and 3.16, both T-BRNN-pre from [75] configuration and Corr-BiRNN from [76] significantly outperformed my lightweight WE-BiLSTM model mainly due to their better performance for commas and question marks. However, it is questionable whether the complex structure of these punctuation recovery systems (including attention mechanism, or joint prediction of punctuation and capitalization) enables them to operate

Table 3.15: Punctuation restoration results for English manual transcripts

Manual Transcript	Model	Comma			Period			Question			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(6,6)	45.6	26.7	33.7	59.4	57.0	58.2	52.4	23.9	32.8	77.2
	WE-BiLSTM	55.5	45.1	49.8	65.9	**75.1**	70.2	57.1	52.2	54.5	59.8
	T-BRNN-pre [75]	**65.5**	47.1	54.8	73.3	72.5	72.9	**70.7**	**63.0**	**66.7**	**49.7**
	Corr-BiRNN [76]	60.9	**52.4**	**56.4**	**75.3**	70.8	**73.0**	70.7	56.9	63.0	50.8
On-line mode	MaxEnt-(10,1)	44.9	23.7	31.0	53.4	50.1	51.7	50.0	21.7	30.8	83.2
	noWE-LSTM	47.3	**42.7**	44.9	60.9	50.4	55.2	**68.2**	32.6	44.1	76.4
	WE-LSTM	**56.3**	40.3	**47.0**	**61.2**	**60.5**	**60.8**	55.5	**43.5**	48.8	**68.1**
	T-LSTM [92]	49.6	41.1	45.1	60.2	53.4	56.6	57.1	43.5	**49.4**	74.0

Table 3.16: Punctuation restoration results for English ASR transcripts

ASR Transcript	Model	Comma			Period			Question			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mode	MaxEnt-(6,6)	40.6	23.9	30.1	56.2	53.5	54.8	31.6	17.1	22.2	84.0
	WE-BiLSTM	46.8	39.6	42.9	60.7	70.3	65.1	44.4	45.7	45.0	72.5
	T-BRNN-pre [75]	**59.6**	42.9	49.9	**70.7**	**72.0**	**71.4**	60.7	48.6	54.0	**57.0**
	Corr-BiRNN [76]	53.5	**52.5**	**53.0**	63.7	68.7	66.2	**66.7**	**50.0**	**57.1**	65.4
On-line mode	MaxEnt-(10,1)	42.6	23.9	30.7	53.2	48.9	51.0	33.3	17.1	23.0	87.0
	noWE-LSTM	40.2	**39.3**	39.7	56.2	46.6	51.0	**76.5**	38.2	**51.0**	86.5
	WE-LSTM	**48.8**	37.1	**42.2**	**57.6**	**57.3**	**57.4**	41.2	41.2	41.2	**78.3**
	T-LSTM [92]	41.8	37.8	39.7	56.4	49.3	52.6	55.6	**42.9**	48.4	83.7

in real-time scenarios. I consider the high recall of periods by my WE-BiLSTM models as a good achievement both on manual and ASR transcripts. In on-line mode, my WE-LSTM system achieved the overall best result. Without using pre-trained word embeddings (noWE-LSTM) my results are getting very close to the T-LSTM configuration.

3.5.5.3 Character-level and Combined Text-based Experiments

Secondly, a character-level embedding CNN-RNN model, and combined word- and character-level models were evaluated in [C9], focusing only on the off-line mode. The neural network based approaches were discussed in Section 3.3 and 3.4. For English models, in the notation of character- and word-level models i stands for the chunk size. The word-level English models were also re-trained from scratch. The hyperparameters of the English models are shown in Table 3.17.

The obtained results are presented in Table 3.18 for the manual transcripts and in Table 3.19 for the automatic (ASR) transcripts, respectively.

Comparing the model types, it is an interesting result that smaller chunk sizes yielded

Table 3.17: Hyperparameters of WE-BiLSTM and CE-CNN-BiLSTM models for English

Language	Model	Chunk Size	Vocab. Size	Embedding dimension	#LSTM cells	Batch size	Optimizer	Filter length	#Filters	Stride	Dilation Rate	Size of MaxPooling Window	Patience
ENG	WE-BiLSTM-200	200	27,244	100	512	128	RMSProp	N/A	N/A	N/A	N/A	N/A	2
	WE-BiLSTM-100	100											
	CE-CNN-BiLSTM-100	100	100	60			Adam with Nesterov Momentum	5	60	1	2	25	4
	CE-CNN-BiLSTM-200	200		70			RMSProp		70		1	20	2

Table 3.18: Punctuation restoration results for English manual transcripts

Model Type	Model name	Comma			Period			Question			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Char-level	(1) CE-CNN-BiLSTM-100	56.0	41.0	47.3	**70.2**	70.6	**70.4**	55.3	56.5	55.9	59.3
	(2) CE-CNN-BiLSTM-200	52.0	35.2	42.0	67.0	66.5	66.7	53.5	50.0	51.7	64.8
Word-level	(3) WE-BiLSTM-100	55.2	44.5	49.3	66.0	68.3	67.1	53.1	37.0	43.6	61.2
	(4) WE-BiLSTM-200	56.3	45.4	50.3	63.4	66.3	64.8	57.5	50.0	53.5	63.1
Hybrid	(1)+(3)	**62.4**	**53.7**	**57.7**	68.4	72.6	**70.4**	64.1	54.3	58.8	**53.0**
	(2)+(4)	61.9	52.2	56.6	64.0	**73.9**	68.6	**66.7**	**65.2**	**65.9**	57.0
Weighted Ensemble	0.4*(1)+0.6*(3)	60.1	43.4	50.4	67.3	70.5	68.9	64.7	47.8	55.0	57.7
	0.31*(2)+0.69*(4)	56.9	41.7	48.1	63.2	67.2	65.1	59.5	47.8	53.0	63.2

better performance in English. Moreover, the results of CE-CNN-BiLSTM-100 character-level model on manual transcripts are also noteworthy for its superior performance compared to my word-level models. Hybrid approaches among combined models give the best result by manual transcripts, improving SER by around 13% and 10% relative to my word-level approaches with smaller (100) and bigger (200) chunk size, respectively. Moreover, the hybrid and weighted ensemble models have a quite similar performance on ASR transcripts, improving SER by 4% relative to my word-level RNN model. Investigating the weights, the moderate dominance of word-level model also appears in the English ensemble models.

3.5.5.4 Textual-acoustic Experiments

Finally, I was highly motivated to adapt my textual-acoustic concept to English language too, and compare my results to the findings of [110, 109]. Please note that the direct

Table 3.19: Punctuation restoration results for English ASR transcripts

Model Type	Model name	Comma			Period			Question			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Char-level	(1) CE-CNN-BiLSTM-100	44.2	35.0	39.1	**65.3**	64.5	64.9	42.1	45.7	43.8	72.9
	(2) CE-CNN-BiLSTM-200	42.1	29.9	35.0	62.5	63.2	62.8	42.5	48.6	45.3	75.8
Word-level	(3) WE-BiLSTM-100	48.3	40.7	44.2	61.9	66.1	63.9	**68.2**	42.9	52.7	71.1
	(4) WE-BiLSTM-200	47.8	40.0	43.6	58.4	62.8	60.5	58.6	48.6	53.1	75.2
Hybrid	(1)+(3)	50.1	**48.2**	**49.1**	63.5	67.1	**65.3**	60.0	42.9	50.0	68.3
	(2)+(4)	50.0	45.7	47.8	58.6	**68.4**	63.1	51.4	**51.4**	51.4	72.5
Weighted Ensemble	0.4*(1)+0.6*(3)	**51.6**	39.8	44.9	63.8	66.3	65.0	66.7	45.7	54.2	**68.2**
	0.31*(2)+0.69*(4)	49.8	38.7	43.6	59.2	64.3	61.6	62.1	**51.4**	**56.2**	73.4

comparison of the results is not applicable, because the dataset of the MGB challenge [154] is not publicly available. I rather focus on the language specificities in general and the role of the embeddings. I applied the transfer learning method [152] to my pre-trained word-level and character-level models (see the hyperparameters in Table 3.20), while the prosodic model learned from scratch, yielding better punctuation performance with a longer (200) chunk size. The English hybrid models are constructed in the same way as described in Section 3.5.3 for Hungarian. The evaluation sets were partly changed, as executing the forced alignment algorithm for one audio recording was unsuccessful, so its manual and ASR transcripts were dropped from the overall set.

Table 3.20: Hyperparameters of the individual model for English

Input	Model	Chunk Size	Vocab. Size	Embedding dimension	#LSTM cells	Batch size	Optimizer	Filter length	#Filters	Stride	Size of MaxPooling Window	Patience
Words	W	200	27,244	100	512	128	RMSProp	N/A	N/A	N/A	N/A	2
Chars	C	200	100	70	512	128	RMSProp	5	70	1	20	2
Prosody	P	200	N/A	N/A	512	16	RMSProp	N/A	N/A	N/A	N/A	3

Table 3.21: Slot Error Rates for English models

Model Type	SER REF	SER ASR
C	70.3	78.2
W	67.5	73.5
P	83.6	83.8
C+W	64.9	71.8
C+P	62.8	67.3
W+P	64.9	67.9
C+W+P	57.8	64.9

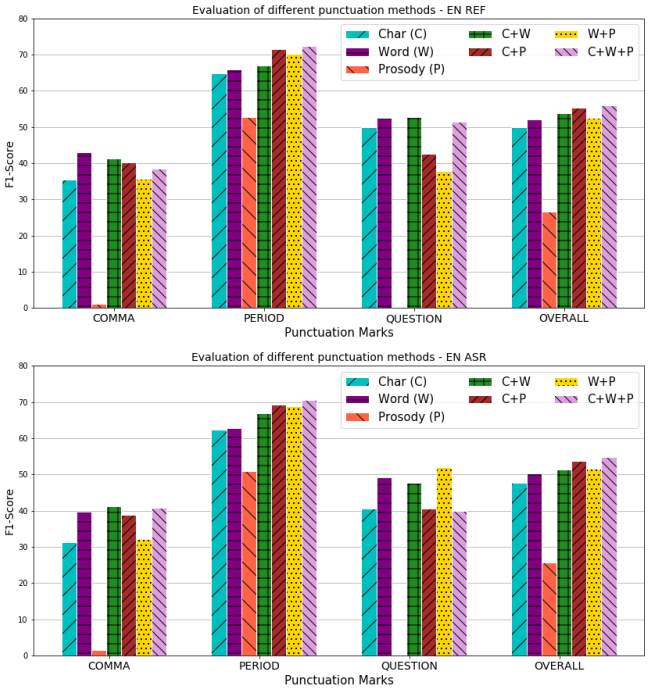

Figure 3.10: F1-results on reference and ASR transcripts in English punctuation, including textual-acoustic models

F1-results for English manual and ASR transcripts are presented in Fig. 3.10 (by WER: 18.7% for ASR), while SER values are shown in Table 3.21.

Considering SER, the C+W+P model yields the best overall performance on reference transcripts ($F1 = 58.8\%; SER = 57.8\%$); this feature triplet lead to a superior performance over W baseline by 14.4% relative, which is significant by $p < 0.05$. Furthermore, also the C+W+P model performs the best on English ASR ($F1 = 56.2\%; SER = 64.9\%$). For English, adding prosody lead to an improvement of 7.6% over the W baseline on ASR transcripts (significant by $p < 0.05$). The English results follow the trends seen for the Hungarian language; when the C+W and also the W+P combination performs superior to the W model, the involvement of all three features leads to the most accurate punctuation restoration. To the best of my knowledge, no prior work is known to use the word, character and prosody feature triplet for English punctuation.

3.5.5.4.1 Comparison of the Experimental Results for Automatic Punctuation between Hungarian and English

Taking a look at the Venn-diagrams in Fig. 3.11, the reference and the ASR cases is also balanced for English like in Hungarian, however, there is less significant gain from adding prosody. The contribution of the character- and word-level features are more emphasized. This can be explained partly by the differences between the two languages.

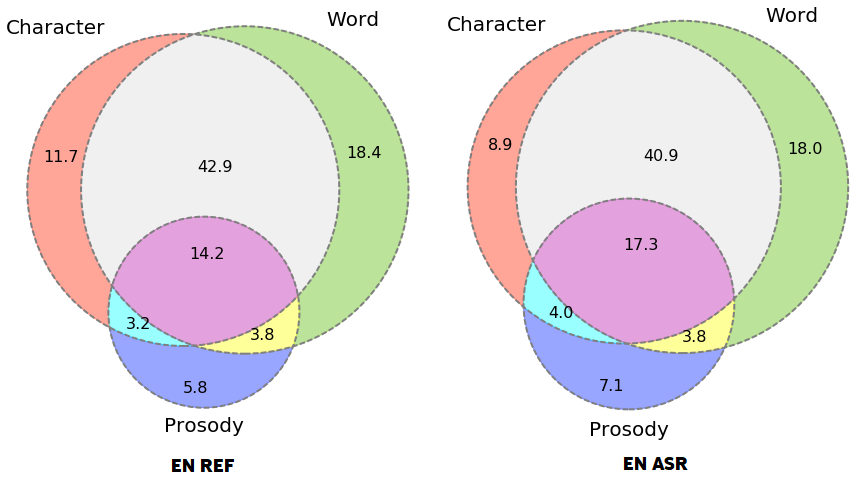

Figure 3.11: Coverage of the English individual models

WER is often higher in agglutinating languages compared to English in corresponding, similar tasks [148], however by this comparison, the filtered sub-corpus of Hungarian Broadcast Dataset (with rather formal genres) has lower overall WER than the English corpus

with more informal TED Talks. In English, periods are efficiently recovered by text-based (character-level and word-level) models, hence prosody contributes only moderately to this punctuation category according to the results depicted in Figure 3.10. However, performance for commas is weaker than for periods for all models, which is opposite to Hungarian. These differences between English and Hungarian can be explained by their fixed vs. relatively free word order and also by the highly agglutinating nature of Hungarian resulting in extreme large vocabularies[4].

Considering the W model, word embeddings operate by capturing the semantic (and also pragmatic) relations of individual words. In a language with constrained word order such as English, the set of words and the role of the words occurring at sentence boundaries is less variable than in Hungarian. Another aspect of the phenomenon results from the fact that word embeddings are less powerful for a language of intense agglutination. The higher number of possible word forms means that the embeddings have more Out-of-Vocabulary (OOV) words (OOV rates are: 6.1% for Hungarian, 4% for English), and their estimation is less robust resulting from the higher overall variability of the data. The latter constitutes the main problem. Using word embeddings enhanced by character N-grams and matching embeddings to the ASR vocabulary can help the OOV problem, and it also improves semantic accuracy as shown for other languages in [155], but not for Hungarian. The unconstrained word order problem still remains an issue, and according to my findings and the observation of the authors of [156], the Hungarian word embeddings enhanced by character N-grams show smaller semantic accuracy measured on word analogy tasks than the English ones [155] despite training them on large corpora. On the other hand, through the C model, character-level information can be exploited with hybrid models, hence I did not use embeddings enhanced by character N-grams in this work.

Seeing higher prediction power for commas in Hungarian is fairly in line with the above hypothesis: if we consider the most common cases where a comma is used – to separate clauses where most often a conjunction word known by the embedding occurs; or in enumerations where typically words with some similarity in the semantic space are involved – embeddings can perform well for these situations.

[4]Agglutinating is linked to free word order too: since case endings define clear grammatical relations within a sentence, the words can be moved around more freely by preserving almost the same core meaning.

3.5.6 Example for Applying Automatic Punctuation for Automatic Summarization

In this subsection, I demonstrate through the example of spoken document summarization how punctuation can be used on top of ASR transcripts: first, a textual transcript is obtained by using the ASR tool; secondly, the text summarization module performs sentence level tokenization on the ASR output. Finally, it applies a ranking of sentences for extractive summary [23], or generates an abstractive summary [157] on the segmented text. In this pipeline, two types of errors can occur: word recognition errors and sentence level tokenization errors. Moreover, both propagate further into the summarization pipeline. Therefore, I evaluated my punctuation approaches with objective measures through the task of extractive summarization of Hungarian spoken documents [J4, C14].

A subset of the Hungarian Broadcast Dataset was used for our[5] experiments, selecting the transcripts of 10 broadcast blocks with overall 500 utterances of 8143 word tokens in total, including weather forecasts, broadcast news and sport news, because these genres were the top most three groups regarding ASR performance, by 6.8%, 10.1%, and 21.4% WER values respectively. Considering manual and ASR transcripts, and manual and automatic punctuation, we created four different types of transcript:

1. Manual transcripts - manual punctuation (MT-MP)
2. ASR transcripts - manual punctuation (AT-MP)
3. Manual transcripts - automatic punctuation (MT-AP)
4. ASR transcripts - automatic punctuation (AT-AP)

The automatic punctuation of these transcripts was done with the word-level bidirectional Recurrent Neural Network (RNN) model presented in Section 3.2. The punctuation marks covered include commas, periods, question marks and exclamation marks. Fig. 3.12 illustrates the number of sentences in the individual blocks.

Switching to automatic punctuation (AP) has obviously some impact on the number of sentence tokens. It is also obvious that the number of sentences is equal for MT-MP and AT-MP, however, the sentence boundaries can be different for the cases MT-AP and AT-AP. Typically, the substitution of a comma by a period, or the insertion of extra periods increase the number of sentences in these transcript types.

[5]These experiments were performed with the contribution of György Szaszák and Valér Kaszás.

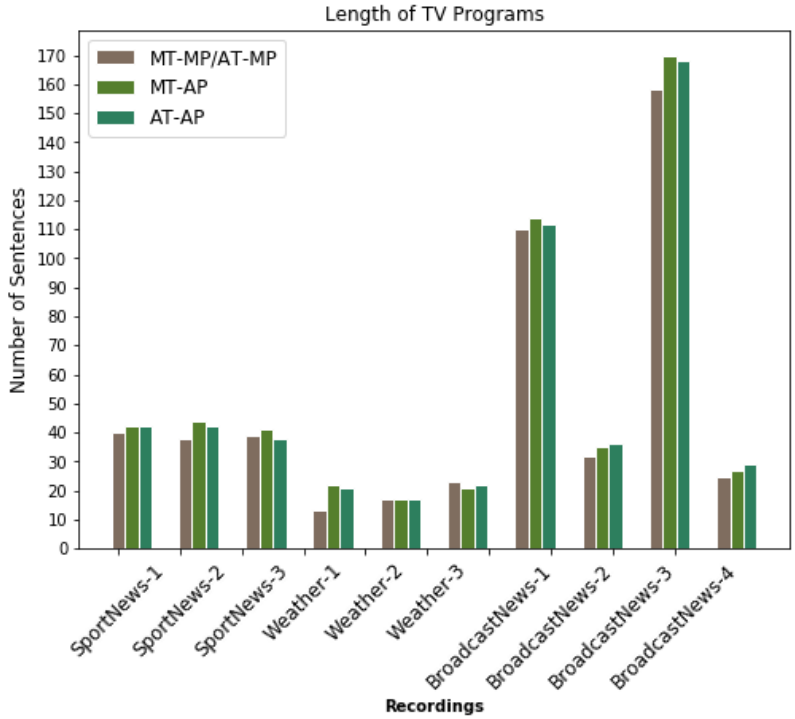

Figure 3.12: The effect of automatic punctuation on the number of sentences per recordings

We used the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric family [158] as evaluation measure, which is commonly used in text summarization. The basic idea behind ROUGE is to compare word overlap between the automatically produced summary and the reference summaries, which may be either human-produced, or derived from highlights (abstracts). In our case, we selected a popular vector space modelling tool, Gensim which uses the BM25 [159] scoring function for sentence ranking to provide automatic extractive summarization, of AT-AP, AT-MP, and MT-AP transcripts, and also for MT-MP for benchmarking. The ratio parameter of this summarizer was set to provide summaries from one-quarter of all sentences of the original transcripts (provide a top N/4 ranking for the document composed of N sentences). These summaries were compared with human-produced references, which were provided by three annotators.

The ROUGE metric family proposes strict measures to assess summaries: counting F1-scores based on recall and precision for a sequence of overlapping words between reference and automatic summaries – known as N-grams – is a common practice, where N is set usually between 1..3. With N-gram metrics, we obtain a strict, but accurate evaluation in terms of coverage between the two summaries, as the word order is taken into account, because it has high impact on meaning of the complete sentence (or summary). We selected four ROUGE-score variants of F1 by our evaluation:

1. ROUGE-1: Unigram-based score of ROUGE
2. ROUGE-2: Bigram-based score of ROUGE
3. ROUGE-L: Longest Common Subsequence (among sequence N-grams)-based score of ROUGE
4. ROUGE-SU4: Skip-Bigram with a maximum skip distance of 4

Figure 3.13. shows the overall ROUGE-scores for the extractive summarization experiments, for the four types of transcripts. Please note that usually the F1-scores are well below 100% for ROUGE, as was proven by [160].

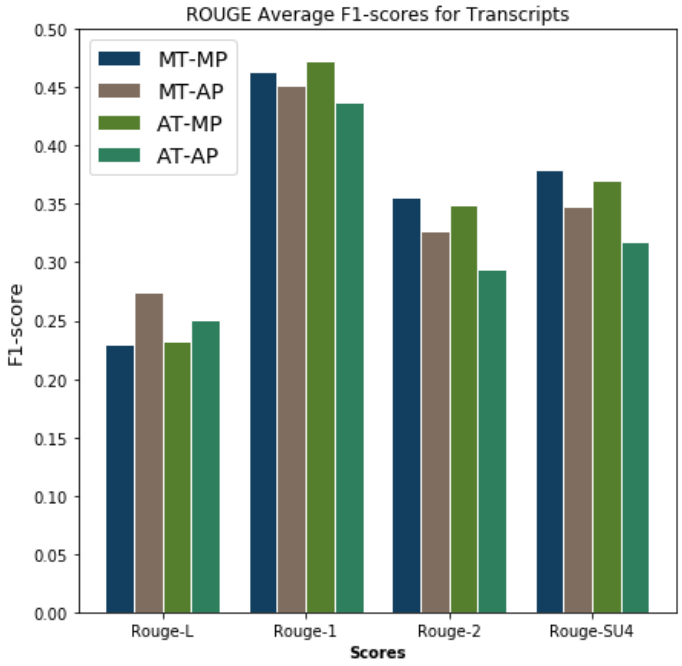

Figure 3.13: Summary of ROUGE-scores for the different speech transcripts

As it is expected, the ROUGE-L shows the lowest values, according to the strictest criterion for N-gram matching, and assuming that the short length of the original documents is also an important factor. Of course, a user-focused evaluation would allow deeper insight into the phenomenon, as outlined by the authors of [161]. On the contrary, the ROUGE-1 provides the highest scores due to its unigram approach. Comparing the texts with automatic punctuation, switching from manual to automatic transcripts results in relative 5-9% performance drop in summarization ROUGE-scores. Except for ROUGE-L, the manually punctuated transcripts performed better than the automatically segmented variants. However, the relative 2-3% differences in performance shows the superiority of MT-MP and

Model Type	ROUGE Score	Punctuation Score	Pearson Correlation (p=0.05)
AT-AP	L	F1 (overall)	0.785
AT-AP	L	Slot Error Rate	-0.786
AT-AP	1	F1 (overall)	0.674
AT-AP	1	Slot Error Rate	-0.635
AT-AP	2	F1 (overall)	0.723
AT-AP	2	Slot Error Rate	-0.687
AT-AP	SU-4	F1 (overall)	0.713
AT-AP	SU-4	Slot Error Rate	-0.672

Table 3.22: Punctuation Score - ROUGE Score Correlations

AT-MP in 2-2 cases.

In general, the results show that the effect of punctuation errors is more significant than the effect of transcription errors on automatic summarization. More precisely, as usually the sentence boundaries count in these tasks, and question and exclamation marks are under-represented in the investigated topics, in this case, the primary effect is derived from the period-related errors. I confirmed my hypotheses with Pearson-correlations between punctuation scores and ROUGE scores at p=0.05 level. The significant results on p=0.05 level are shown in Table 3.22. According to the correlation values of AT-AP category, the interplay of punctuation and transcription errors is highly pronounced. However, I could not confirm significance for MT-AP category, but I also experienced the dominant effect of punctuation errors on ROUGE-scores there.

This way, I investigated the effects of punctuation errors (combined with transcription errors) on an SLU-related task, such as automatic summarization. I showed that the automatically punctuated texts yield fairly comparable results to the reference transcripts. I argue that the capabilities of Hungarian ASR-systems extended with automatic punctuation post-processing module can be useful for automatic summarization methods.

In this chapter, I proposed different architectures for punctuation restoration in Hungarian speech transcripts (both manual and ASR-produced closed captioning texts). Besides the "single input" (word- and character-based) models, combined approaches were also presented and evaluated with objective metrics. I also made an objective evaluation for my word-level approach through an SLU-related task, namely automatic summarization. In the next chapter, I show the detailed subjective evaluation of the automatic punctuation, involving native Hungarian end-users.

Chapter 4

User-centric Evaluation of Automatic Punctuation

4.1 Background

4.1.1 The Importance of User-centric Evaluation

The aim of closed captioning (CC) technology is to provide textual information for people on a visual display, for example for broadcast media services. Part of the audience using CC is composed of deaf and hard of hearing (DHH) people, so CC is expected to make them TV programs also accessible. How this can be obtained w.r.t. the presentation method (captioning style) and understandability of the content (reliability) has been declared by legislation and investigated by several studies. The author of [162] presents laws of broadcast media accessibility by legislative bodies and implementation of CC in USA, Canada, United Kingdom (as the leader of CC technology in Europe), and Brazil. In most of the cases, the captions are specialized subtitles for DHH people; it provides not the simple, raw transcription of speech but also paralinguistic information which may not be perceived by DHH audience (for example when a phone is ringing). Moreover, the amount and the placement of displayed words shall be aligned to the human reading capabilities, otherwise it will be hard to track and comprehend information. Of course, the automation in CC is also important, because the manual preparation may be time-consuming. The authors of [163] applied automatic summarization which displays closed captions at a readable speed. However, they did not replace the re-writing with automated solution (i.e. created the closed captions by hand), and the summarization may cause uncertainty for hearing impaired people, when the mouth movement is really different from the caption. Hong et al. created

dynamic captioning which includes automatic face recognition and script-to-face mapping, in order to better recognize the speakers, moreover, the variation of voice volume is also illustrated [164]. What is more important to us, that nowadays we can find more examples where the ASR technology appears in CC systems such as [165, 166], even for Hungarian [12]. As it is expected, the accuracy and latency issues are highly emphasized by the DHH audience. In fact, the authors of [166] held a co-workshop together with them, proposing some important requirements regarding real-time CC.

From the perspective of CC and spoken language understanding (SLU), traditional ASR performance indicators such as WER may not be the optimal choice, as WER correlates with the human subjective performance poorly. Thus, new metrics were proposed and the effects of different errors on the users' understandability were thoroughly examined to point out possible future improvements [167, 168, 169, 170, 171, 172]. It is common practice to ask subjects to answer questions about the content of a given text, to get feedback about the understandability, when reference and ASR transcripts are compared [173]. Human Perceived Accuracy (HPA) [167] intended to measure the semantic similarity/difference between these transcripts, relying on the saliency of the words using Inverse Document Frequency (IDF) and weighting the different ASR-errors (insertion, substitution and deletion); it was evaluated with the help of hearing people. Nevertheless, human perception can be different for hearing people and DHH audience, however, there was less attention paid to the latter group. The authors of [171] analyzed the effect of ASR errors on text understandability for DHH people. They categorized 10 types of errors and measured their correlation with comprehension scores; the finding was that errors connected to the word length, the word importance (measured with TF-IDF) and the Part-of-Speech tags have emphasized role in the understanding. The same authors developed the Automated-Caption Evaluation (ACE) metric aligned to the subjective evaluation of DHH people [172]. ACE consists of two main components; a word predictability score to identify keywords in a text and semantic distance as an approximation of the deviation in meaning due to errors, to predict the degree of usability of the automatically generated captions by DHH users. The subjective judgements demonstrated much higher correlation with ACE than with WER.

Inserting punctuation marks automatically is one of the possibilities to make the CC systems more intelligible and provide user-friendly captions [174]. As I mentioned earlier, the aim of punctuation is to help human reading and understanding, both for people with normal hearing, and also for DHH people. The previous examples showed that it is not easy to separate readability and understanding; punctuation marks can also lead to better

readability, which is a condition for understanding a text easily, i.e. understanding is the final purpose of reading. During the subjective test procedure to be presented hereafter, I also received feedback that these two are felt closely related in perception.

To the best of my knowledge, no previous studies considered subjective evaluation for the DHH group regarding punctuated captions, and I am also unaware of systematic subjective assessment of punctuation in ASR (also w.r.t. its objective evaluation metric, Slot Error Rate). The next subsections present an evaluation study in Hungarian to cover this gap; both with transcripts for healthy people, and with subtitled TV-programs for DHH people.

4.2 User-centric Evaluation of the Word-based RNN Model

4.2.1 Research Questions

Starting from the point of closed captioning, the goal is to transfer to the users in writing what is meant in speech. This is implemented by using an ASR to caption spoken content and convert it into a written form. My interest is condensed into the following questions:

- $Q1$: To what extent is the understandability and the readability of the written form influenced by the existence or lack of accurate punctuation marks?
- $Q2$: How does the automatic placement of punctuation affect people's subjective feelings? Can we point out any important difference compared to manual punctuations?
- $Q3$: How does the difficulty of ASR task affect the subjective judgement of automatically punctuated texts?
- $Q4$: How strong is the correlation between subjective preference and objective error rates (WER, SER)?

4.2.2 Research Materials

To make the assessment of the ASR possible with the punctuation model and separately, I consider the following caption text types (in brackets I provide abbreviation used hereafter to refer to the text type used in the captions):

1. Manual transcripts - manually added punctuation (MT-MP)
2. ASR transcripts - manually added punctuation (AT-MP)

3. Manual transcripts - automatic (RNN-based) punctuation (MT-AP)

4. ASR transcripts - automatic (RNN-based) punctuation (AT-AP)

5. Manual transcripts - no punctuation (MT-NP)

6. ASR transcripts - no punctuation (AT-NP)

I carried out the subjective evaluation with the word-based RNN automatic punctuation model (see Section 3.2.4). The caption text types were selected from the Hungarian Test Dataset, balanced for the genres (BC-News, BC-Conversations, Magazines, Sport Magazines, Sport News and Weather Forecasts). Altogether 36 test sessions were constructed of texts of short passages or part of interviews with up to 10 sentences each, prepared in the 6 combinations outlined above (manual and ASR transcripts with no, manual and automatic punctuation). The covered punctuation marks include commas, periods, question marks and exclamation marks. Colons and semicolons are mapped to commas, all other punctuation marks are removed.

A large group of hearing people evaluated caption texts, while the DHH people evaluated real captions shown together with video. Further details of study setups and their results are provided in Sections 4.2.3 and 4.2.4.

4.2.3 Subjective Test with Normal Hearing Subjects

Thesis III.A. [C11] *I confirmed with statistical experiments, that the ASR output enriched by my automatic punctuation approach is significantly more understandable for the normal hearing subjects than the unpunctuated transcript.*

The subjects were instructed to read a text and rate it on a scale as follows:

- Excellent (Grammar (word or punctuation) errors were not perceptible, text is well understood)
- Good (Some grammar errors were perceptible, but understood)
- Fair (Texts should be read more times due to the grammar errors, but the whole content can be interpreted finally)
- Poor (Only some specific part of the whole text is understandable due to the grammar errors)

- Bad (It's impossible to understand the content due to the grammar errors)

My first interest is to confirm that the understandability and/or readability of texts is influenced by the existence or lack of punctuation marks. My hypothesis is that punctuation marks can significantly help human reading and understanding. I am also interested in comparing the manual and the automatic methods for punctuation in order to qualify the proposed punctuation model, i.e. whether a non-perfect punctuation is rather helpful or disturbing. By using the punctuation model on ASR output, it is assumed that ASR errors are more disturbing than punctuation errors and hence, it is needed to confirm that punctuation still adds a benefit in readability when used on such output.

The subjective test was organized involving 181 participants (age: $\mu = 28.23$ and $\sigma = 9.20$), including 121 men and 60 women, leading to 460 ratings overall. Each subject rated at least 2 and at most 14 text snippets. All subjects were native Hungarian speakers.

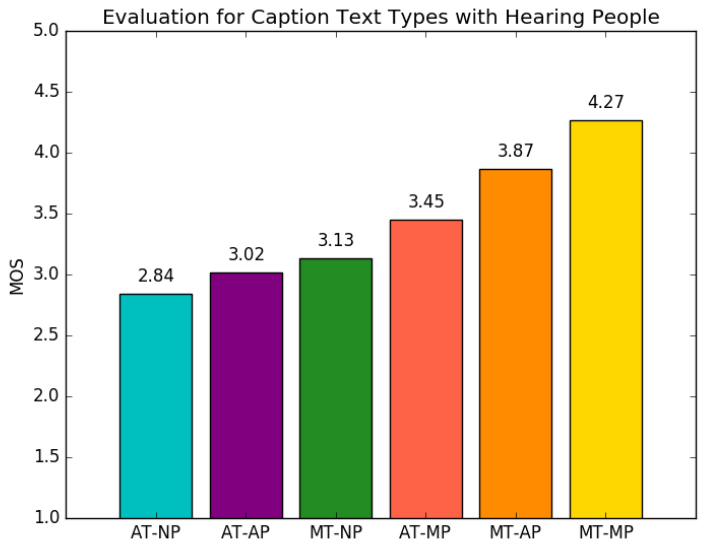

Figure 4.1: MOS for the 6 caption text types

Mean Opinion Scores (MOS) computed on the ratings are shown in Fig. 4.1. In addition, Table 4.1 shows the outcome of pairwise Mann-Whitney U-tests (p=0.05), which are used to test whether the MOS obtained for the different forms of the texts are significantly different. ANOVA was not applicable for my analysis, because the normality of the data cannot be assumed, and the sample sizes are small(N<=30) in many cases.

Table 4.1: Results of pairwise Mann Whitney U-test,
* marks significant MOS difference ($p<0.05$), U-values in the brackets

Caption Text Types	MT-MP	MT–AP	AT-MP	MT-NP	AT-AP	AT-NP	Mean
MT-MP	1						**4.27**
MT-AP	**0.002* (1874.5)**	1					**3.87**
AT-MP	**0* (1683.5)**	**0.007* (2318)**	1				**3.45**
MT-NP	**0* (1000.5)**	**0* (1435)**	**0.017* (2525)**	1			**3.13**
AT-AP	**0* (1286)**	**0* (1725.5)**	**0.013* (2771.5)**	0.436 (2896.5)	1		**3.02**
AT-NP	**0* (731.0)**	**0* (1063.5)**	**0* (2014.5)**	**0.033* (2247)**	0.376 (2828)	1	**2.84**

4.2.3.1 Importance of Manual Punctuation (AD Q1)

Taking a look at the MOS values it is not surprising, that the highest MOS is associated with MT-MP and the lowest MOS with AT-NP. For manual transcripts, the manual punctuation was significantly more preferred by the test subjects over the unpunctuated MT-NP (see also MT-NP row in Table 4.1). In case of ASR transcripts, MOS of manually punctuated captions (AT-MP) is significantly superior not only to AT-AP and AT-NP but also to MT-NP, which means that even if the captions contain word errors, the presence of precise punctuation can counteract this and leads to better understandability. These findings suggest that **people have a clear preference for punctuated texts**.

4.2.3.2 Subjective Impression on Automatic Punctuation (AD Q2)

Although for manually transcribed captions the highest MOS was measured with manual punctuation (comparing MT-MP to MT-AP and MT-NP), scores also showed significant preference for automatically punctuated samples (MT-AP) over unpunctuated ones (MT-NP). Considering the captions with ASR transcription, AT-AP and AT-NP, which basically constitute the two alternative use-cases available during automatic closed captioning with or without automatic punctuation, a MOS superior for AT-AP is visible, but this difference is not significant. However, here the effects (and errors) of automatic transcription and punctuation interplay. Based on the fact that MT-AP is significantly better than AT-MP I hypothesize that word errors due to ASR have greater impact on subjective opinions than errors introduced by automatic punctuation. I investigate this issue further in the next subsection.

4.2.3.3 Genre Analysis for Caption Text Types (AD Q3)

Fig. 4.2 shows the MOS values for each of the 6 genres (see also Section 3.2.1) revealing additional notes about the 6 rated caption text types.

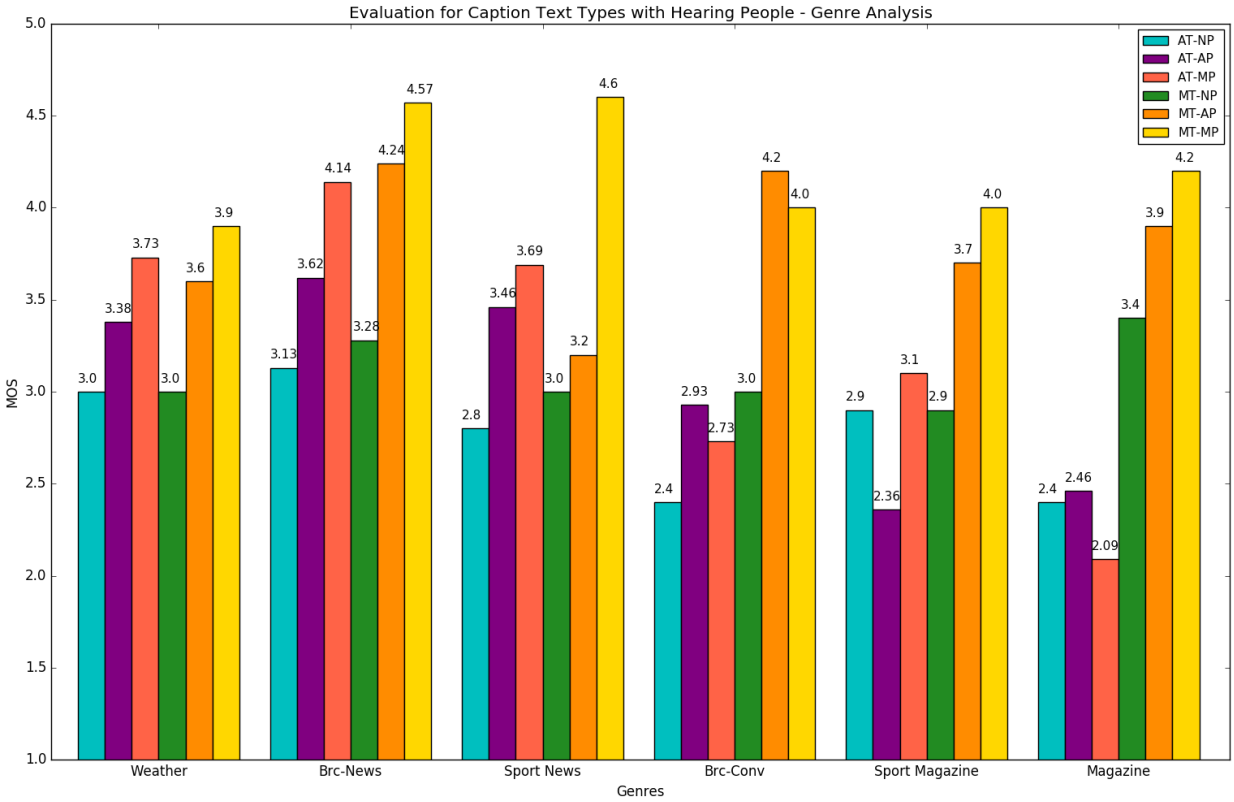

Figure 4.2: MOS for the 6 caption text types - Genre Analysis

It can be observed that for certain genres, such as weather forecasts, broadcast news, or sport news, WER is lower, whereas closed captioning for broadcast conversations, sport magazines, magazines, operates at higher WER. For genres characterized by higher WER, all caption text types based on ASR transcripts (AT-MP, AT-AP, AT-NP) have considerably lower MOS than their counterparts on manual transcripts, and the unpunctuated captions are not perceived as significantly better than the automatically punctuated (RNN) alternative.

At high WER the punctuation may become senseless once the word sequence is heavily altered. As this obviously has an impact on punctuations, i.e. if words are misrecognized around punctuation slots, the original sense of punctuation gets lost; and as word errors are more disturbing than punctuation errors (this follows from the MOS values in Fig. 4.1), word errors can make a caption uninterpretable irrespective of whether it is punctuated or not.

Re-running the evaluation for data with **WER limited at around 20%**[1] (with the top most three groups), **I obtained a significant difference between the MOS of AT-AP and AT-NP groups** (3.51 vs. 3.02, respectively; p=0.016). These are good news regarding the helpfulness of automatic punctuation. Beyond this threshold, MOS for the automatically punctuated AT-AP version remains higher, but the difference to AT-NP is not significant. It can be concluded that hearing people accept automatic punctuations if the quality of the ASR transcription reaches a certain accuracy level, considering the effect of word errors.

Therefore, for broadcast news, sport news and weather forecasts, automatic punctuation of ASR captions seems worth to apply in a real-time environment and it is significantly better in terms of MOS than the unpunctuated version, where the predictability of the text is better. Findings here also reveal that the more spontaneous the speech gets (e.g. more fillers and disfluencies come up), the more word errors occur. This way, the less trivial is to tell (even for human annotators, proven by Inter-Annotator Agreement-experiments) where sentences end, as spontaneous speech often shows irregular sentence patterns [175, 176].

4.2.3.4 The Relationship between Subjective and Objective Metrics (AD Q4)

After my mainly subjective observations in Section 4.2.3.1 and 4.2.3.2, I analyze the role of WER and SER by measurable means. Whether MOS depends on the WER of ASR and the SER of the punctuation model seem to be obvious, but the strength of the correlation between MOS and the error measures used to qualify ASR and punctuation is an interesting question. Therefore, I calculate Pearson's Correlation Coefficient (PCC) and Spearman's Rho Correlation Coefficient (SRC) to get an insight into this issue. Computing these values will allow to evaluate to some extent the contribution of word errors and punctuation errors to the overall subjective ratings (i.e. to MOS).

Fig. 4.3 shows MOS as a function of SER for the different genres, whereas Table 4.2 shows the PCC and the SRC between MOS and SER / F1.

Table 4.2: SER - MOS and F1 - MOS correlation for MT-AP category

Caption Text Type	Correlated pairs	PCC (p-value)	SRC (p-value)
MT-AP	**SER-MOS**	-0.393 (0.001)	-0.365 (0.002)
MT-AP	**F1-MOS**	0.405 (0.001)	0.388 (0.001)

[1] As said, such WERs for the highly agglutinating Hungarian may be regarded to constitute roughly the upper limit of real examples within the investigated tasks.

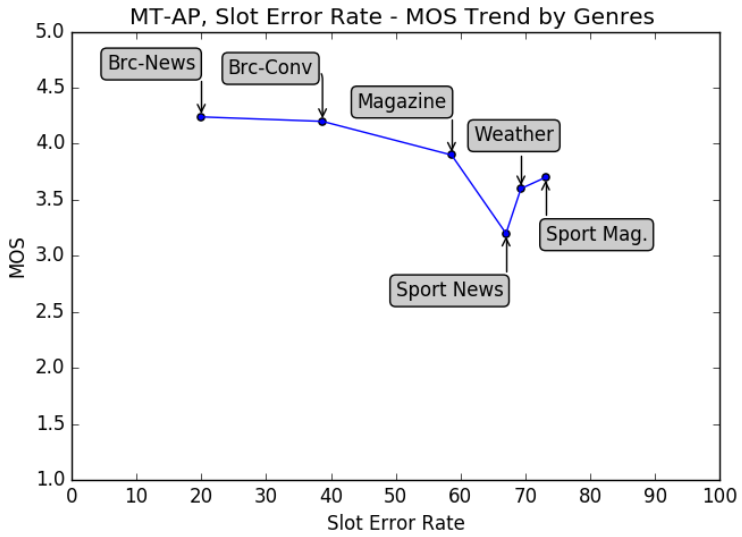

Figure 4.3: MOS as a function of SER for MT-AP caption text type

Albeit this analysis is noisy in terms of its genre (i.e. speech style) dependency, this also means that if word errors are not present (by MT-AP Caption Text Type), the punctuation influences the subjective ratings rather weakly.

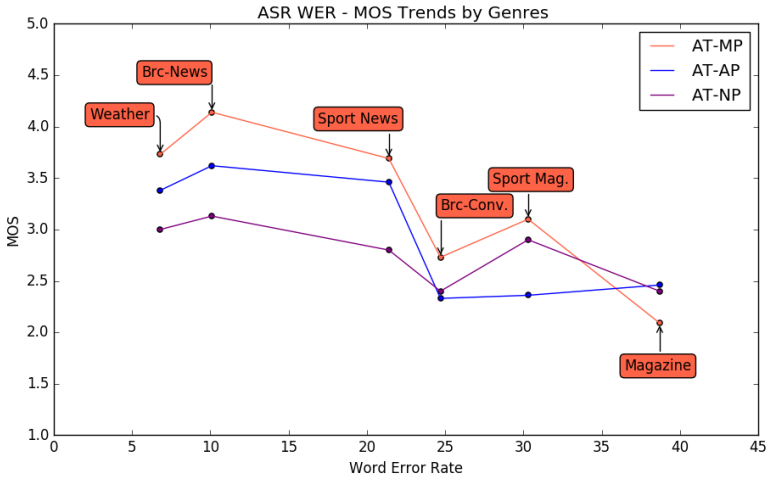

Figure 4.4: WER - MOS trends for ASR-based caption text types

Fig. 4.4 shows MOS as a function of WER for ASR caption text types. Correlation between WER and MOS increases as punctuation is improved (see Table 4.3). I explain this as follows: MOS is a subjective aggregate of word and punctuation accuracies (and other factors such as familiarity with topic, etc.). If punctuation accuracy gets higher, WER and MOS become more correlated, as the contribution of WER to MOS is more significant. This phenomenon shows us how important punctuation is, even if there was no

overall significant difference confirmed in MOS between AT-AP and AT-NP in section 4.2.1. However, correlation remains weak, which raises the mentioned issues regarding the limits of appropriateness of the objective measures, when we expect it to reflect subjective impression.

Table 4.3: Various correlation pairs for ASR-based caption text types

Caption Text Types	Correlated pairs	PCC (p-value)	SRC (p-value)
AT-NP	**WER-MOS**	-0.259 (0.026)	-0.259 (0.026)
AT-AP	**WER-MOS**	-0.389 (0.000)	-0.408 (0.000)
AT-MP	**WER-MOS**	-0.639 (0.000)	-0.580 (0.000)
AT-AP	**WER-SER**	0.522 (0.000)	0.530 (0.000)
AT-AP	**SER-MOS**	-0.123 (0.270)	-0.203 (0.066)
AT-AP	**F1-MOS**	0.201 (0.068)	0.203 (0.066)

Fig. 4.5 shows MOS as a function of SER. At the first glance it becomes obvious that word errors have a greater impact on MOS and that no consistent relationship can be confirmed between MOS and SER, except for probably the MT-AP case, when WER=0%. However, switching to the AT-AP strategy, which is the realistic use case, this correlation could not be confirmed any more.

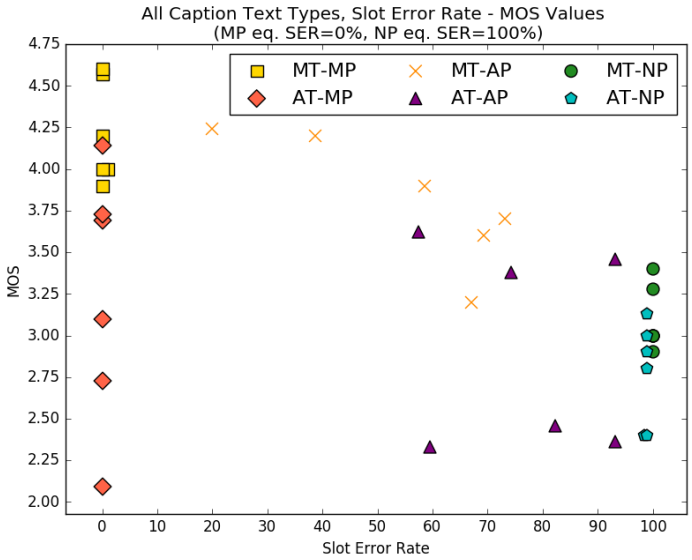

Figure 4.5: SER-MOS plots for all caption text types

I attribute this to the dominance of word errors on one hand, and to the impact of speech style on the other hand, but any of these two does not explain sufficiently such a weak correlation. Therefore, by acknowledging the limits of the evaluation presented here, one raises the question whether SER at all is a suitable measure in terms of reflecting subjective impression. A similar behaviour is seen with the F1-MOS relationship, with PCC a bit higher, but still showing a weak correlation which I attribute again to the higher impact of word errors, speech style dependency and mismatch between the objective and subjective measures.

Formal (higher MOS: weather forecast, BC and sport news) and spontaneous (lower MOS: BC and sport conversations, magazines) speaking styles are separated into the two observed clusters in Fig. 4.5. When both word and punctuation errors are present, SER was not informative at all regarding user rating.

4.2.3.4.1 GAM Approach

We [2] experienced, that the user score depends on many factors. Supposing that some of these are determined by word and punctuation errors, a Generalized Additive Model (GAM) [177] can help in identifying the ratio by which such factors X_i contribute to the user score Y. With GAM, the user score can be decomposed as follows:

$$g(E[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_n(x_n), \tag{4.1}$$

where $g(.)$ is the link function and $E[.]$ gives the expectation.

Defining $X_1 ... X_n$ such that they represent insertion, substitution and deletion errors for words and for punctuation marks ($n = 6$) using smoothing spline estimates for $f_i(x_i)$, it turns out that punctuation insertion and substitution errors alone with the number of punctuation slots explain 32.3% of the variance observed in MOS. Deletion errors in punctuation were rare and hence we could not determine their contribution to MOS with sufficient certainty. Nevertheless, the higher impact of insertion errors in punctuation coincides with intuition, i.e. insertion errors were expected to be more disturbing as they provide a false structuring of the information, likely to counteract grammatical rules and constraints, whereas a deletion may be easier to recover by humans.

[2] This experiment was performed with the contribution of András Beke.

4.2.4 Subjective Tests with DHH Subjects

Thesis III.B. [C11] *I confirmed with statistical experiments, that the ASR-based closed captions enriched by my automatic punctuation approach are more understandable for Deaf and Hard-of-Hearing (DHH) end-users than the unpunctuated transcripts. DHH subjects rated the automatically punctuated manual and ASR transcripts to be significantly more understandable.*

Thesis III.C. [C11] *I confirmed with statistical experiments, that DHH people are not able to perceive significant difference between the manual and the automatic punctuation; these methods contributed to the comprehensibility in a similar manner.*

Previously I showed that hearing people tend to prefer punctuated captions, moreover, they also accept automatic punctuation if the quality of the ASR transcription exceeds a certain level, considering the effect of word errors. Closed captioning – although, if translated, it may be useful also for language learners not speaking the language of audio – is primarily required for DHH people, as European directives also oblige television companies to provide subtitles for all broadcast media. Thus I carried out subjective tests with DHH subjects to see how my former findings generalize to the primary audience of closed captioning.

In order to get an authentic picture, I preferred a test setup as close to real usage conditions as possible: 18 DHH students (aged 13-14 years) were asked to view short, 1-1.5 minute long coherent, muted and subtitled video recordings in a classroom experiment [3]. I simulated static captioning, which means that the whole subtitle block is shown at once (one-shot appearance). I wanted to avoid possible understandability difficulties arising from unfamiliarity with the genres (e.g. broadcast conversation about economics, politics), moreover, as the effect of WER was examined thoroughly in the previous studies, I took these two factors into consideration by selection, to have more focus on the punctuation errors, thus I selected weather forecast and sport news samples.

Each video snippet was prepared with the same 6 captioning strategies, but instead of direct scoring, a comparative assessment was carried out. The students were divided into 3 smaller disjunct groups of 6 students each; they watched the samples pairwise for the same video snippet with different caption texts, and were then asked to indicate which version of the two they preferred. Each group saw 4 video snippets, ensuring that each student saw all the 6 caption text types. Figs 4.6 and 4.7 show two examples of automatically punctuated ASR closed captions.

[3] The recordings were provided by Speechtex Ltd.; their support is greatly appreciated

The main goal of the experiment was to decide whether punctuation marks in the captioning can help the DHH people in understanding the content, while focusing also on, whether they perceive any changes in the quality of texts and punctuation. After watching subsequently the same video snippet with two alternative subtitles, subjects performed comparisons with the help of a prepared drawing, referring to a scale, with indication to draw the arrow of the scale proportional to their subjective impression in favour of one of the alternatives (see Fig. 4.8 as an example).

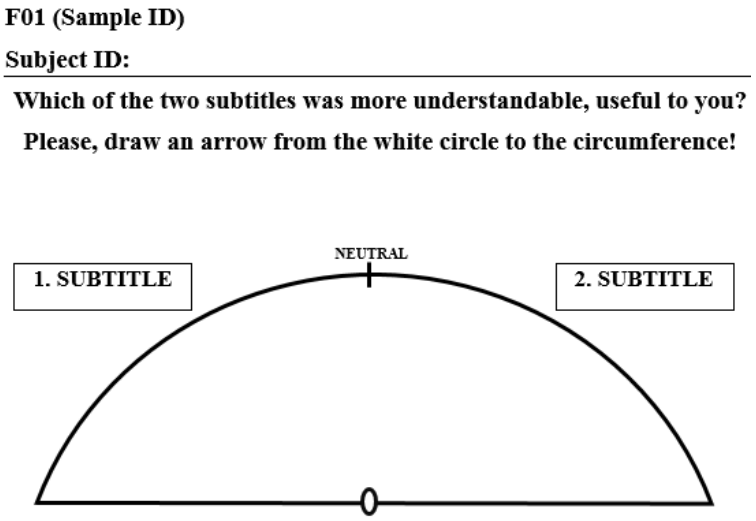

Figure 4.8: Evaluation Sheet(translated to English)

Finally, this was quantized to three grades representing preference for either of the samples or a neutral opinion.

4.2.5 DHH-audience-related Experimental Results

Table 4.4. summarizes results (votes received) on manual transcripts comparing pairwise the different punctuation strategies, whereas Table 4.5. contains the results obtained using ASR captions and, again, different punctuation strategies. Table 4.6. contains the overall results of the two experiments.

Analyzing these **results shows a clear preference for punctuated captions** (MP vs. NP, AP vs. NP), **with an interesting, albeit not significant superiority of automatic punctuation over the manual one** (AP vs. MP). The differences are more pronounced in videos related to sport news compared to weather forecasts. Results on ASR and on manual transcripts show the same tendencies.

Figure 4.6: A Hungarian subtitle for sport news: "fordulás után már kevésbé forogtak veszélyben a kapott gól már nem született. [As there wasn't any dangerous situation by received, the result was not changed.]"–> In this case, there is one mismatched word (kapuk <–> the goals), hence, there is a missing period before *gól*, as the word context is changed around it.

Figure 4.7: A Hungarian subtitle for weather forecast: "..és 20-22 fokig melegszik a levegő, tehát húsokat nem is érzékelnek majd a frontból. [The temperature can get as high as 20-22 Celsius, so they won't feel meat from the front.]" –> In this case, there are two mismatched words (túl sokat <–> too much), but the two punctuation marks are correct.

Table 4.4: Closed Captioning results for manual transcripts

Genres	MT-AP vs. MT-NP			MT-AP vs. MT-MP			MT-MP vs. MT-NP		
	MT-AP	Same	MT-NP	MT-AP	Same	MT-MP	MT-MP	Same	MT-NP
Weather	2	3	1	2	3	1	1	3	2
Sport news	5	1	0	1	4	1	3	3	0
All (%)	58.3	33.3	8.3	25.0	58.3	16.7	33.3	50.0	16.7

Table 4.5: Closed Captioning results for ASR transcripts

Genres	AT-AP vs. AT-NP			AT-AP vs. AT-MP			AT-MP vs. AT-NP		
	AT-AP	Same	AT-NP	AT-AP	Same	AT-MP	AT-MP	Same	AT-NP
Weather	1	4	1	2	2	2	3	2	1
Sport news	4	0	2	2	4	0	5	1	0
All (%)	41.7	33.3	25.0	33.3	50.0	16.7	66.7	25.0	8.3

Table 4.6: Joint (MT+AT) Closed Captioning results

Genres	AP vs. NP			AP vs. MP			MP vs. NP		
	AP	Same	NP	AP	Same	MP	MP	Same	NP
Weather	3	7	2	4	5	3	4	5	3
Sport news	9	1	2	3	8	1	8	4	0
All (%)	50.0	33.3	16.7	29.2	54.2	16.7	50.0	37.5	12.5

Pairwise exact tests [178] were performed, although in several cases the number of votes was not sufficient to find significant differences at p=0.05. Nevertheless, some significant differences were confirmed, hence I highlight only on the significant observations:

1. For ASR Captions (both sport news and weather forecasts), there is a significant difference between the votes for captions with manually added punctuation marks (MP) and captions without punctuation marks (NP), favouring the punctuated one (p=0.048).

2. Overall for all relevant pairs, there is a significant number of ratings (50% of the cases, 24 from 48), preferring the punctuated subtitles (MT-MP, MT-AP, AT-AP, AT-MP) versus the lack of punctuation (AT-NP+MT-NP) (p=0.012).

3. For sport news, there is a significant difference between the votes for captions with automatically restored punctuation marks (MT-AP+AT-AP) and captions without punctuation marks (MT-NP+AT-NP), favouring the punctuated one (p=0.012).

4. For sport news, subjects were unable to make a difference between manual (MT-MP, AT-MP) and RNN punctuation (MT-AP+AT-AP); the number of votes reflecting neutral opinion on the difference is significantly higher than the two others (p=0.048).

Considering the results for weather forecasts, the picture is not that clear, as I mostly failed to confirm significance of the differences; I suppose two reasons behind this: (i) the higher ratio of unfamiliar, genre-specific words require extra effort by young people for understanding; (ii) the fast speech tempo, which may cause that there is no sufficient time available to read the consecutive caption blocks. This, indeed, could be alleviated only by a careful design of the captioning style (i.e. using a kind of fast abstractive summarization or selection of relevant group of words), which was outside of the scope of my experiments.

Examining the votes person by person, 61% of them (11/18) have a positive balance in favour of the enhanced captions (despite of some votes for unpunctuated subtitles), which means DHH people preferred punctuations in videos.

Chapter 5

Conclusions and the Summary of the Theses

During my research, I focused on the structural segmentation of ASR-produced speech transcripts, as the segmented input is the key component of the further processing regarding Spoken Language Understanding (SLU). I presented the two main approaches for segmentation of the speech. The first one is to mark phrase boundaries based on prosodic features, which are sentence-like, or smaller units, but the main point is that they often correlate well with the syntax of the language. The second one is, which can be more familiar for people who are non-experts in natural language processing and speech technology; the automatic restoration of punctuation marks into ASR output, or punctuation for short.

I proposed both prosody- and text-driven solutions; the main focus was on the development for my native language, Hungarian, but I considered the adaptability for foreign languages as well. I organized my results into three Thesis Groups; in this chapter, I give a short summary for each thesis within them.

Thesis Group I.: Atom Decomposition-based Prosodic Stress Detection and Phonological Phrasing

In my first thesis group, I proposed a prosody-based segmentation technique for the fixed stress Hungarian and adapted the method for French. Before the discussion of my results, I provided some information about the importance of prosody w.r.t. its relation to the information structure, and its role in current SLU-tasks. I provided a detailed technical overview of stress detection for different foreign languages (including the pool of the used prosodic features), and an overview of automatic phrasing methods, finally highlighting the solutions

to the languages of my particular interest. For both Hungarian and French, I used an intonation modelling approach called WCAD, as a base method, exploiting features for stress detection (and hence phonological phrasing) directly.

As my stress detection approach is evaluated only indirectly, through the task of automatic phonological phrasing task, I didn't formulate theses regarding the efficiency of stress detection separately. Nevertheless, I formulated auxiliary hypotheses aligned to the stress characteristics of the Hungarian and French, in correlation with the places of the extracted atoms, after they were assigned to the specific syllables of the word. After the confirmation of these hypotheses, I implemented my automatic phrasing methods. First, I formulated theses for Hungarian language, as follows:

Thesis I.A. [C3, J2] *I experimentally confirmed, that my atom decomposition-based phonological phrasing method significantly outperforms the HMM/GMM baseline method (by relative 7% in F1, on the investigated Hungarian corpus).*

With this thesis, I proved that my WCAD-based solution can be as efficient for a fixed-stressed language in terms of PP-detection as an earlier HMM/GMM-method, which applies supervised learning. In the contrary, as I presented, the construction of PP-segmentation with my method relies on some major characteristic of Hungarian. As my solution uses syllable-level information for the segmentation, it is obviously more accurate in the segmentation than the baseline method concerning timestamps of phrase onsets and ends.

To reach further improvement in phrase-level segmentation for Hungarian, I combined my solution with the baseline method; as the result of successful aggregation, I formulated the following thesis:

Thesis I.B. [C2, J2] *Combining my atom decomposition-based solution and the HMM/GMM baseline approach, the obtained hybrid model yields a significant increase in the performance of automatic phrasing over the HMM/GMM baseline (by relative 11% in F1, on the investigated Hungarian corpus).*

This hybrid solution establishes a common PP-segmentation based on the boundaries received from two sources, while considers the elimination of possible overlapping intervals. This approach exceeded the performance of both base methods in F1-scores.

After the analysis of French stress characteristics (which showed somewhat smaller correlation with atom-syllable pairs compared to Hungarian findings), I also proposed a WCAD-based PP segmentation model adaptation for this language. I formulated my thesis for French language:

Thesis I.C. [J2] *I have experimentally confirmed, that the adaptation of my atom decomposition-based phonological phrasing method for French yields comparable results to the baseline HMM/GMM approach, while combining the two methods significantly outperforms the baseline (by relative 6% in F1, on the investigated corpus).*

With this result, I showed that as the step of atom extraction is language-independent (it relies on the probability of voicing and the energy features), a successful model adaptation can be possible across fixed-stressed languages to some extent; this way, further languages can be investigated in the future. It is important to mention that compared to the baseline HMM/GMM approach, my approach requires the transcript of the speech recording, either a gold reference or ASR hypothesis. However, it may seem a huge constraint, at least, the ASR transcript can be easily produced, so my solution is also applicable in practice. Moreover, with the help of this transcript, a lexical stress detection task can be investigated in the future.

Thesis Group II.: Automatic Punctuation with Neural Networks

In my second thesis group, I proposed automatic punctuation approaches for Hungarian, using neural network models, which is the state-of-the-art methodology for this field. First, I proved the necessity of the punctuation insertion into ASR output with speech technology-related examples, then I offered a survey of punctuation paradigms, from the very first language model-related solutions, finishing with the current, Deep Learning-based structures, where the sequence-to-sequence labelling RNN models and the Neural Machine Translation alike encode-decoder approaches are 'competing' with each other. I reviewed the pros and cons of text-based and prosody-based solutions as well.

For Hungarian, as far as I know, there was only a prosody-based automatic punctuation approach, when I started to design text-based models. The selected baseline approach for my experiment was a Maximum Entropy-based universal sequential tagger called Huntag. I adapted Huntag to deal with my classification task, where I considered the prediction of four punctuation marks (commas, periods, question marks and exclamations marks), and blank (O) label, as the 5^{th} class. I mainly worked with broadcast data, as the original idea was to

integrate my punctuation module to an ASR-based closed captioning (CC) system.

In my first experiments, regarding CC, I implemented on-line solutions and compared their performance degradation to off-line solutions, where the past and additionally, the future context of a given word was also considered. I compared the performance of my word-level Recurrent Neural Network-based models with MaxEnt solutions by manual and ASR transcripts, and formulated the following thesis:

Thesis II.A. [C6, C7, C17] *I have experimentally confirmed, that my word-based RNN model for automatic punctuation significantly outperforms the Maximum Entropy-based baseline system, in off-line operation mode, for Hungarian language (20% relative improvement on reference transcripts, and 10% relative improvement on ASR transcripts in SER).*

The results showed not only the efficiency of RNN over the MaxEnt approach, but the general capability of dealing with the agglutinative, morphologically rich and free word order nature of Hungarian language for this task. To compare with the prosody-based segmentation in Thesis Group 1, Hungarian has some fixed characteristics for comma placement, but it is usually not the case for other punctuation marks, also reflected by the results.

As the word-level solution was constrained to a fixed-size vocabulary (to limit the computational complexity of the model) and pre-trained embeddings (which usually incorporate millions of word token), I proposed a character-level off-line approach. In this model, the key role belongs to the 1D-Convolution–MaxPooling feature extractor layer pair, besides that, the sequence-to-sequence tagging capability is still guaranteed by a recurrent layer. It is important that there is no constraint for the alphanumeric characters (i.e. there isn't any missed out element), every character is represented in the semantically connected embedding vector space (which is learned by the model itself from the initialization state). I formulated the following thesis for this model:

Thesis II.B. [C9] *I have experimentally confirmed, that automatic punctuation using a character-based CNN-RNN model for Hungarian is also possible; my CNN-RNN model yields slightly lower performance compared to my word-based punctuation approach presented in Thesis II.A.*

I concluded, that the corresponding results convey two important messages; the importance of the word forms by predicting punctuation mark for sentence ends, and in parallel, the

strength of low-level (character-level) features, due to the elimination of the aforementioned constraints.

At this point, I realized the potential of developing hybrid approaches, involving multiple features; first, I proposed a model with character-word feature pair. However, the most fruitful decision was to create models including feature triplets with the involvement of prosody, and even the character-prosody feature pair yielded better results in the automatic punctuation compared to the word-level baseline. Besides the Hungarian experiments, I performed analyses for English models, and compared the differences and the similarities of these two languages concerning automatic punctuation. I formulated one thesis for these experiments:

Thesis II.C. [C9] *I have experimentally confirmed, that my hybrid punctuation model consisting of character- and word-based components significantly outperforms my word-based punctuation model for Hungarian in terms of SER (by 3-9% relative for manual transcripts and by 2-3% relative for ASR transcripts on the investigated corpus).*

Thesis II.D. [C13, C18] *I have experimentally confirmed, that my hybrid punctuation model consisting of character-, word- and prosody-based components significantly outperforms my character-word based hybrid on the manual and ASR transcripts of the involved Hungarian corpus (by 8% and 6% relative in SER, respectively).*

Albeit I investigated many variants of recurrent neural network architectures for the task of automatic punctuation, this research topic can be also potentially extended for a long-term; for example, trying more sophisticated word embeddings (including sub-words, creating ELMo [179] or BERT [180] representations), or using acoustical embeddings for Hungarian, even making experiments with other DNN-architectures (e.g. encoder-decoder) as well.

Thesis Group III.: User-centric Evaluation of Automatically Punctuated ASR-transcripts

Finally, I presented the results of the user-focused evaluation of my automatic punctuation method. It is obvious, that the manual (accurate) punctuation yields better understandability for the people than reading raw, unpunctuated texts; the most important question was that what can be concluded regarding automatic punctuation, especially if it is applied to ASR transcripts. I made a full evaluation involving six caption text types, combining the modes of transcript and the punctuation methods in all possible ways. Both normal hearing people and the primary audience of ASR-closed captioning, a group of deaf and

hard-of-hearing subjects evaluated the texts in different test procedures. Normal hearing subjects read texts, and based on their ratings, Mean Opinion Scores were calculated for the six transcripts. After statistical pairwise comparisons, I formulated the following thesis:

Thesis III.A. [C11] *I confirmed with statistical experiments, that the ASR output enriched by my automatic punctuation approach is significantly more understandable for the normal hearing subjects than the unpunctuated transcript.*

This thesis was determined besides a certain margin of automatic transcription errors. Obviously, if the ASR Errors propagate into the post-processing step of automatic punctuation, due to the word errors, the syntactic-semantic changes cause uncertainty in the role of punctuation. However, the role of punctuation marks will be irrelevant for the users, as the perception of transcription errors can result in syntactically or semantically erroneous, hence less understandable sentences. As a future perspective, the different types of transcription and punctuation errors can be combined to a common weighted metric, as I showed, that the currently used objective metrics, Slot Error Rate and Word Error Rate weakly correlate with human ratings.

I organized a classroom experiments with a DHH-audience, showing short video samples with different subtitles, including automatically punctuated versions. Based on their ratings, I also performed statistical analysis, then I formulated two theses;

Thesis III.B. [C11] *I confirmed with statistical experiments, that the ASR-based closed captions enriched by my automatic punctuation approach are more understandable for Deaf and Hard-of-Hearing (DHH) end-users than the unpunctuated transcripts. DHH subjects rated the automatically punctuated manual and ASR transcripts to be significantly more understandable.*

Thesis III.C. [C11] *I confirmed with statistical experiments, that DHH people are not able to perceive significant difference between the manual and the automatic punctuation; these methods contributed to the comprehensibility in a similar manner.*

During the experiments, I realized that introducing proof-of-concepts, or enhancing new applications for DHH-audience requires special attention, i.e. besides the quality of the transcription, the way of visualization also needs careful design; the involvement of a wider audience would be fruitful for further experiments.

Chapter 6

Applicability of My Results

Using my practical approaches, the functional expansion of the raw ASR-output is applicable, relying on both prosodic and textual information, which is an important step for supporting Spoken Language Understanding for Hungarian as well. I proposed model adaptations proving that my solutions can be extended to other languages.

Results of Thesis Group I. can be used for stress detection and phonological phrasing, which yield an acoustically coherent speech segmentation. Phonological phrases (PPs) can serve as input for a document summarization system, but not exclusively; I also showed that with some high-level features derived from PPs (duration and word count), diagnostic classification support can be developed for pathological subjects [C5, C16].

Results of Thesis Group II. allow for the automatic placement of punctuation marks in the ASR output. As an off-line solution, punctuation restoration can be performed on fully generated ASR-transcripts; in this way, a wide range of natural language processing related tasks are allowed to be executed, beginning with the syntactic and semantic (dependency) parsing, but also generating question-answer pairs on the basis of punctuated transcripts. Considering my on-line solution with low latency, it can be used in ASR-based closed captioning systems to improve understandability of television programs.

In Thesis Group III. I demonstrated a proof-of-concept application with automatic punctuation. The need for punctuation was clearly confirmed by the end-users (especially by DHH-audience); to keep the low latency operation, this development expects the close collaboration of speech technology experts, the end-users and the broadcast data providers.

As a future goal, a hybrid document summarization approach can be designed, for example by a weighted combination of prosodic stress prediction and syntactic-semantic information extraction from the structured text with automatically restored punctuation marks.

Bibliography

[1] A. Turing, "Can digital computers think?(1951)," *B. Jack Copeland*, p. 476, 2004.

[2] W. Ward, "The CMU air travel information service: Understanding spontaneous speech," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[3] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding." in *Interspeech*, 2013, pp. 3771–3775.

[4] S. L. Tóth, D. Sztahó, and K. Vicsi, "Speech emotion perception by human and machine," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer, 2008, pp. 213–224.

[5] K. Vicsi and G. Szaszák, "Using prosody to improve automatic speech recognition," *Speech Communication*, vol. 52, no. 5, pp. 413–426, 2010.

[6] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[9] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*. IEEE, 2013, pp. 6645–6649.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[11] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," *arXiv preprint arXiv:1802.08395*, 2018.

[12] Á. Varga, B. Tarján, Z. Tobler, G. Szaszák, T. Fegyó, C. Bordás, and P. Mihajlik, "Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach," in *Proceedings of SPECOM*. Springer, 2015, pp. 105–112.

[13] G. Gosztolya, T. Grósz, and L. Tóth, "GMM-Free Flat Start Sequence-Discriminative DNN Training," in *Interspeech 2016*, 2016, pp. 3409–3413. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-391

[14] J. Hajič, "Morphological tagging: Data vs. dictionaries," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 2000, pp. 94–101.

[15] L. Laki, "Investigating the Possibilities of Using SMT for Text Annotation," in *1st Symposium on Languages, Applications and Technologies*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2012.

[16] G. Orosz and A. Novák, "Purepos 2.0: a hybrid tool for morphological disambiguation," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, 2013, pp. 539–545.

[17] K. Pajkossy and A. Zséder, "The hunvec framework for NN-CRF-based sequential tagging," in *LREC*, 2016.

[18] K. Kann, J. Bjerva, I. Augenstein, B. Plank, and A. Søgaard, "Character-level Supervision for Low-resource POS Tagging," in *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, 2018, pp. 1–11.

[19] G. Recski and D. Varga, "A Hungarian NP chunker," *The Odd Yearbook*, vol. 8, pp. 87–93, 2009.

[20] T. Váradi, E. Simon, B. Sass, M. Gerőcs, I. Mittelholtz, A. Novák, B. Indig, G. Prószéky, and V. Vincze, "Az e-magyar digitális nyelvfeldolgozó rendszer," 2017.

[21] G. Szaszak and A. Beke, "Using phonological phrase segmentation to improve automatic keyword spotting for the highly agglutinating Hungarian language," Idiap, Tech. Rep., 2013.

[22] A. Beke and G. Szaszák, "Combining NLP techniques and acoustic analysis for semantic focus detection in speech," in *5th IEEE Conference on Cognitive Infocommunications (CogInfoCom 2014)*. IEEE, 2014, pp. 493–497.

[23] A. Beke and G. Szaszák, "Automatic summarization of highly spontaneous speech," in *International Conference on Speech and Computer*. Springer, 2016, pp. 140–147.

[24] A. Langus, E. Marchetto, R. A. H. Bion, and M. Nespor, "Can prosody be used to discover hierarchical structure in continuous speech?" *Journal of Memory and Language*, vol. 66, no. 1, pp. 285–306, 2012.

[25] M. Selting, "On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation," *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, vol. 6, no. 3, pp. 371–388, 1996.

[26] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, pp. 485–946, 1998.

[27] L. Czap and J. M. Pintér, "Intensity feature for speech stress detection," in *Carpathian Control Conference (ICCC), 2015 16th International*. IEEE, 2015, pp. 91–94.

[28] D. Sportiche, H. Koopman, and E. Stabler, *An introduction to syntactic analysis and theory*. John Wiley & Sons, 2013.

[29] L. Hunyadi, *Hungarian sentence prosody and universal grammar: on the phonology-syntax interface*. Lang, Peter, Gmbh, Internationaler Verlag Der Wissenschaften, 2002, vol. 13.

[30] E. Selkirk, "The syntax-phonology interface," in *International Encyclopaedia of the Social and Behavioural Sciences*. Oxford: Pergamon, 2001, pp. 15 407–15 412.

[31] M. Nespor and I. Vogel, *Prosodic phonology: with a new foreword.* Walter de Gruyter, 2007, vol. 28.

[32] W. J. Levelt, *Speaking: From intention to articulation.* MIT press, 1993, vol. 1.

[33] A. Christophe, S. Peperkamp, C. Pallier, E. Block, and J. Mehler, "Phonological phrase boundaries constrain lexical access i. adult data," *Journal of memory and language*, vol. 51, no. 4, pp. 523–547, 2004.

[34] H. Truckenbrodt, "On the relation between syntactic phrases and phonological phrases," *Linguistic inquiry*, vol. 30, no. 2, pp. 219–255, 1999.

[35] D. Hakkani-Tür, G. Tür, A. Stolcke, and E. Shriberg, "Combining words and prosody for information extraction from speech," in *Proceedings of Eurospeech 1999*, 1999.

[36] F. Gallwitz, H. Niemann, E. Nöth, and W. Warnke, "Integrated recognition of words and prosodic phrase boundaries," *Speech Communication*, vol. 36, no. 1-2, pp. 81–95, 2002.

[37] C. Lai and S. Renals, "Incorporating lexical and prosodic information at different levels for meeting summarization," in *Proceedings of Interspeech 2014*, 2014.

[38] M. Horne, P. Hansson, G. Bruce, and J. Frid, "Prosodic correlates of information structure in Swedish human-human dialogues," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[39] R. Silipo and F. Crestani, "Prosodic stress and topic detection in spoken sentences," in *String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on.* IEEE, 2000, pp. 243–252.

[40] D. B. Fry, "Experiments in the perception of stress," *Language and speech*, vol. 1, no. 2, pp. 126–152, 1958.

[41] P. Roach, "On the distinction between 'stress-timed' and 'syllable-timed' languages," *Linguistic controversies*, vol. 73, p. 79, 1982.

[42] M. Ponomareva, K. Milintsevich, E. Chernyak, and A. Starostin, "Automated Word Stress Detection in Russian," in *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 2017, pp. 31–35.

[43] M. A. Shahin, J. Epps, and B. Ahmed, "Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning," in *INTERSPEECH*, 2016, pp. 175–179.

[44] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks," *Speech Communication*, vol. 96, pp. 28–36, 2018.

[45] V. Apopei and O. Păduraru, "Towards prosodic phrasing of spontaneous and reading speech for Romanian corpora," in *Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on.* IEEE, 2015, pp. 1–4.

[46] N. Warner, L. Butler, and T. Arai, "Intonation as a speech segmentation cue: Effects of speech style," in *9th Conference on Laboratory Phonology.* Citeseer, 2004, pp. 37–42.

[47] S.-A. Jun, "Prosodic typology," in *The phonology of intonation and phrasing.* Oxford University Press, 2006, pp. 430–459.

[48] L. Varga, *Intonation and stress: evidence from Hungarian.* Springer, 2002.

[49] A. Heba, T. Pellegrini, T. Jorquera, R. André-Obrecht, and J.-P. Lorré, "Lexical Emphasis Detection in Spoken French Using F-BANKs and Neural Networks," in *International Conference on Statistical Language and Speech Processing.* Springer, 2017, pp. 241–249.

[50] P. Martin, "Automatic phrasing in French," in *Proc. 9th International Conference on Speech Prosody 2018*, 2018, pp. 607–611.

[51] G. Christodoulides, A. C. Simon, and I. Didirková, "Perception of Prosodic Boundaries by Naïve and Expert Listeners in French Modelling and Automatic Annotation," in *Proc. 9th International Conference on Speech Prosody 2018*, 2018, pp. 641–645.

[52] B. Gerazov, P.-E. Honnet, A. Gjoreski, and P. N. Garner, "Weighted correlation based atom decomposition intonation modelling," in *Proceedings of Interspeech*, Dresden, Germany, September 2015, pp. 1601–1605.

[53] G. Olaszy, G. Németh, P. Olaszi, G. Kiss, C. Zainkó, and G. Gordos, "Profivox:A Hungarian text-to-speech system for telecommunications applications," *International Journal of Speech Technology*, vol. 3, no. 3-4, pp. 201–215, 2000.

[54] P.-E. Honnet, B. Gerazov, A. Gjoreski, and P. N. Garner, "Intonation modelling using a muscle model and perceptually weighted matching pursuit," *Speech Communication*, vol. 97, pp. 81–93, 2018.

[55] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Second international conference on spoken language processing*, 1992.

[56] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, 2009, pp. 81–84.

[57] B. Gyuris, "Sentence-types, discourse particles and intonation in Hungarian," in *Proceedings of Sinn und Bedeutung*, vol. 13, 2009, pp. 157–170.

[58] K. Mády and F. Kleber, "Variation of pitch accent patterns in Hungarian," in *Proceedings of Speech Prosody 2010*, 2010.

[59] P. S. Roach, S. Amfield, W. Bany, J. Baltova, M. Boldea, A. Fourcin, W. Goner, R. Gubrynowicz, E. Hallum, L. Lamep, K. Marasek, A. Marchal, E. Meiste, and K. Vicsi, "BABEL: An Eastern European Multi-language database," in *International Conf. on Speech and Language*, 1996, pp. 1033–1036.

[60] J.-P. Goldman, P.-E. Honnet, R. Clark, P. N. Garner, M. Ivanova, A. Lazaridis, H. Liang, T. Macedo, B. Pfister, M. S. Ribeiro, E. Wehrli, and J. Yamagishi, "The SIWIS database: a multilingual speech database with acted emphasis," in *Proceedings Interspeech*, 2016, pp. 1532–1535.

[61] G. Szaszák and A. Beke, "Exploiting prosody for automatic syntactic phrase boundary detection in speech," *Journal of Language Modeling*, vol. 0, no. 1, pp. 143–172, 2012.

[62] M. D. Tyler and A. Cutler, "Cross-language differences in cue use for speech segmentation," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 367–376, 2009.

[63] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals,*

Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on. IEEE, 1993, pp. 40–44.

[64] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[65] D. J. Hermes, "Measuring the perceptual similarity of pitch contours," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, pp. 73–82, February 1998.

[66] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 2513–2517.

[67] Y. Marchand, C. R. Adsett, and R. I. Damper, "Automatic syllabification in English: A comparison of different algorithms," *Language and Speech*, vol. 52, no. 1, pp. 1–27, 2009.

[68] G. Szaszák, M. G. Tulics, and M. Tündik, "Analyzing F0 discontinuity for speech prosody enhancement," *Acta Univ. Sapientiae Elect. Mech. Eng*, vol. 6, no. 1, pp. 59–67, 2014.

[69] G. Szaszák and A. Beke, "An empirical approach for comparing syntax and prosody driven stress marking," *The Phonetican Journal of the International Society of Phonetic Sciences*, no. 114, pp. 46–57, 2017.

[70] F. Batista, "Recovering Capitalization and Punctuation Marks on Speech Transcriptions," Ph.D. dissertation, Instituto Superior Técnico, 2011.

[71] A. Novák and B. Siklósi, "Automatic diacritics restoration for hungarian," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2286–2291.

[72] B. Siklósi, A. Novák, and G. Prószéky, "Context-aware correction of spelling errors in hungarian medical documents," *Computer Speech & Language*, vol. 35, pp. 219–233, 2016.

[73] B. Keszler, "Punctuation and interdisciplinarity," *Magyar Nyelv (Hungarian Language)*, vol. 1, no. 103, pp. 1–16, 2007.

[74] W. Gale and S. Parthasarathy, "Experiments in character-level neural network models for punctuation," *Proc. Interspeech 2017*, pp. 2794–2798, 2017.

[75] O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," in *Proceedings of Interspeech*, 2016, pp. 3047–3051.

[76] V. Pahuja, A. Laha, S. Mirkin, V. Raykar, L. Kotlerman, and G. Lev, "Joint Learning of Correlated Sequence Labelling Tasks Using Bidirectional Recurrent Neural Networks," *arXiv preprint arXiv:1703.04650*, 2017.

[77] J. Kolář, J. Švec, and J. Psutka, "Automatic punctuation annotation in Czech broadcast news speech," *SPECOM´ 2004*, 2004.

[78] J. Kolář and L. Lamel, "Development and evaluation of automatic punctuation for French and English speech-to-text," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[79] A. Vāravs and A. Salimbajevs, "Restoring punctuation and capitalization using transformer models," in *International Conference on Statistical Language and Speech Processing*. Springer, 2018, pp. 91–102.

[80] A. Caranica, H. Cucu, A. Buzo, and C. Burileanu, "Capitalization and punctuation restoration for romanian language," *University Politehnica of Bucharest Scientific Bulletin*, 2015.

[81] J. Llombart, A. Miguel, A. Ortega, and E. Lleida, "Wide residual networks 1d for automatic text punctuation," *Proc. IberSPEECH 2018*, pp. 296–300, 2018.

[82] Q. Chen, J. Wu, and T. Su, "Investigating context influence in character Level LSTM methods for Japanese Auto Punctuation," 2018.

[83] A. Moró and G. Szaszák, "A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery," in *Proceedings of Interspeech*, 2017.

[84] C. J. Chen, "Speech recognition with automatic punctuation," in *Proceedings of Eurospeech*, 1999.

[85] E. Shriberg, A. Stolcke, and D. Baron, "Can Prosody Aid the Automatic Processing of Multi-Party Meetings? Evidence from Predicting Punctuation, Disfluencies, and Overlapping Speech," in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[86] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 474–485, 2012.

[87] A. Stolcke, E. Shriberg, R. A. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *ICSLP*, vol. 2, 1998, pp. 2247–2250.

[88] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: A lightweight punctuation annotation system for speech," in *Proceedings of ICASSP*. IEEE, 1998, pp. 689–692.

[89] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *Proceedings of ICASSP*. IEEE, 2009, pp. 4741–4744.

[90] W. Lu and H. T. Ng, "Better punctuation prediction with dynamic conditional random fields," in *Proceedings of EMNLP*. ACL, 2010, pp. 177–186.

[91] E. Cho, J. Niehues, K. Kilgour, and A. Waibel, "Punctuation insertion for real-time spoken language translation," in *Proceedings of the Eleventh International Workshop on Spoken Language Translation*, 2015.

[92] O. Tilk and T. Alumäe, "LSTM for punctuation restoration in speech transcripts," in *Proceedings of Interspeech*, 2015, pp. 683–687.

[93] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech." in *Proceedings of Interspeech*, 2002, pp. 917–920.

[94] J. Driesen, A. Birch, S. Grimsey, S. Safarfashandi, J. Gauthier, M. Simpson, and S. Renals, "Automated production of true-cased punctuated subtitles for weather and news broadcasts," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[95] J. Yi, J. Tao, Z. Wen, and Y. Li, "Distilling knowledge from an ensemble of models for punctuation prediction," *Proc. Interspeech 2017*, pp. 2779–2783, 2017.

[96] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *Proceedings of EMNLP*, 2014, pp. 1532–1543.

[97] A. Öktem, M. Farrús, and L. Wanner, "Attentional Parallel RNNs for Generating Punctuation in Transcribed Speech," in *International Conference on Statistical Language and Speech Processing*. Springer, 2017, pp. 131–142.

[98] C. Hardmeier, "A neural model for part-of-speech tagging in historical texts," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 922–931.

[99] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models." in *AAAI*, 2016, pp. 2741–2749.

[100] D. Klein, J. Smarr, H. Nguyen, and C. D. Manning, "Named entity recognition with character-level models," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 180–183.

[101] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *arXiv preprint arXiv:1511.08308*, 2015.

[102] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[103] J. Chung, K. Cho, and Y. Bengio, "A character-level decoder without explicit segmentation for neural machine translation," *arXiv preprint arXiv:1603.06147*, 2016.

[104] C. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69–78.

[105] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition." in *Interspeech*, vol. 2013, 2013, pp. 1173–5.

[106] L. Tóth, "Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 190–194.

[107] P. Żelasko, P. Szymański, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, "Punctuation prediction model for conversational speech," *Proc. Interspeech 2018*, pp. 2633–2637, 2018.

[108] M. Ballesteros and L. Wanner, "A neural network architecture for multilingual punctuation generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1048–1053.

[109] O. Klejch, P. Bell, and S. Renals, "Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches." in *SLT*, 2016, pp. 433–440.

[110] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on.* IEEE, 2017, pp. 5700–5704.

[111] E. Cho, J. Niehues, and A. Waibel, "NMT-based Segmentation and Punctuation Insertion for Real-time Spoken Language Translation," *Proc. Interspeech 2017*, pp. 2645–2649, 2017.

[112] F. Chollet, "Keras: Theano-based deep learning library," *Code: https://github.com/fchollet. Documentation: http://keras.io*, 2015.

[113] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[114] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[115] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[116] P. J. Werbos *et al.*, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[117] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[118] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[119] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[120] N. Srivastava, "Improving neural networks with dropout," *University of Toronto*, vol. 182, p. 566, 2013.

[121] Y. Bengio, P. Simard, P. Frasconi, *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[122] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[123] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[124] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[125] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[126] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[127] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[128] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 1555–1565.

[129] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Part-of-speech tagging with bidirectional long short-term memory recurrent neural network," *arXiv preprint arXiv:1510.06168*, 2015.

[130] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[131] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[132] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[133] C. N. d. Santos and V. Guimaraes, "Boosting named entity recognition with neural character embeddings," *arXiv preprint arXiv:1505.05008*, 2015.

[134] M. Ballesteros, C. Dyer, and N. A. Smith, "Improved transition-based parsing by modeling characters instead of words with lstms," *arXiv preprint arXiv:1508.00657*, 2015.

[135] F. J. Huang, Y.-L. Boureau, Y. LeCun, *et al.*, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[136] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[137] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *International conference on artificial neural networks*. Springer, 2010, pp. 92–101.

[138] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[139] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[140] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[141] A. Ratnaparkhi *et al.*, "A maximum entropy model for part-of-speech tagging," in *Proceedings of EMNLP*, 1996, pp. 133–142.

[142] M. Makrai, "Filtering Wiktionary triangles by linear mapping between distributed models," in *Proceedings of LREC*, 2016, pp. 2776–2770.

[143] T. Váradi, "The hungarian national corpus," in *In Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas*, 2002, pp. 385–389.

[144] P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, and V. Trón, "Creating open language resources for hungarian," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, C. N, Ed., 2004, pp. 203–210. [Online]. Available: http://eprints.sztaki.hu/7896/

[145] X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation Prediction for Unsegmented Transcript Based on Word Vector," in *Proceedings of LREC*, 2016, pp. 654–658.

[146] A. Stolcke, "SRILM:an extensible language modeling toolkit," in *Proceedings International Conference on Spoken Language Processing*, Denver, US, 2002, pp. 901–904.

[147] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*. IEEE, 2011, pp. 1–4.

[148] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* -. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 487–494.

[149] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of DARPA broadcast news workshop*, 1999, pp. 249–252.

[150] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[151] C. Teleki, S. Velkei, S. L. Tóth, and K. Vicsi, "Development and evaluation of a Hungarian Broadcast News Database," in *Forum Acusticum*, 2005.

[152] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[153] M. Federico, S. Stüker, L. Bentivogli, M. Paul, M. Cettolo, T. Herrmann, J. Niehues, and G. Moretti, "The IWSLT 2011 evaluation campaign on automatic talk translation," in *International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 3543–3550.

[154] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, *et al.*, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 687–693.

[155] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[156] B. Döbrössy, M. Makrai, B. Tarján, and G. Szaszák, "Investigating sub-word embedding strategies for the morphologically rich and free phrase-order hungarian," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 2019, pp. 187–193.

[157] P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 354–358.

[158] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[159] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *arXiv preprint arXiv:1602.03606*, 2016.

[160] N. Schluter, "The limits of automatic summarisation according to rouge," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 41–45.

[161] F. Liu and Y. Liu, "Correlation between rouge and human evaluation of extractive meeting summaries," in *Proceedings of the 46th annual meeting of the association for*

computational linguistics on human language technologies: Short papers. Association for Computational Linguistics, 2008, pp. 201–204.

[162] G. Toledo, "Subtitles for the deaf and hard-of-hearing: Comparing legislation and official orientation for SDH in Brazil and in other countries," *Transletters. International Journal of Translation and Interpreting*, vol. 1, pp. 143–166, 2018.

[163] T. Monma, E. Sawamura, T. Fukushima, I. Maruyama, T. Ehara, and K. Shirai, "Automatic closed-caption production system on TV programs for hearing-impaired people," *Systems and Computers in Japan*, vol. 34, no. 13, pp. 71–82, 2003.

[164] R. Hong, M. Wang, M. Xu, S. Yan, and T.-S. Chua, "Dynamic captioning: video accessibility enhancement for hearing impairment," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 421–430.

[165] M. Federico and M. Furini, "Enhancing learning accessibility through fully automatic captioning," in *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*. ACM, 2012, p. 40.

[166] S. Kawas, G. Karalis, T. Wen, and R. E. Ladner, "Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students," in *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 2016, pp. 15–23.

[167] T. Mishra, A. Ljolje, and M. Gilbert, "Predicting Human Perceived Accuracy of ASR Systems," in *Proc. INTERSPEECH*, 2011, pp. 1945–1948.

[168] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova, D. Ochei, G. Penn, S. Tratz, *et al.*, "Automatic human utility evaluation of ASR systems: does WER really predict performance?" in *INTERSPEECH*, 2013, pp. 3463–3467.

[169] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Eighth International Conference on Spoken Language Processing*, 2004.

[170] N. Itoh, G. Kurata, R. Tachibana, and M. Nishimura, "A metric for evaluating speech recognizer output based on human-perception model," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[171] S. Kafle and M. Huenerfauth, "Effect of Speech Recognition Errors on Text Understandability for People who are Deaf or Hard of Hearing," in *Proc. SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies*, 2016, pp. 20–25.

[172] S. Kafle and M. Huenerfauth, "Evaluating the Usability of Automatically Generated Captions for People who are Deaf or Hard of Hearing," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 2017, pp. 165–174.

[173] D. A. Jones *et al.*, "Measuring the readability of automatic speech-to-text transcripts," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[174] C. E. Johnson, C. S. Antunes, R. S. Zimmerman, J. E. Barron, J. Miller, and A. Khesin, "Intelligent caption systems and methods," Apr. 25 2017, US Patent 9,632,997.

[175] V. Cabarrão, H. Moniz, F. Batista, R. Ribeiro, N. J. Mamede, H. Meinedo, I. Trancoso, A. I. Mata, and D. M. de Matos, "Revising the annotation of a Broadcast News corpus: a linguistic approach." in *LREC*, 2014, pp. 3908–3913.

[176] M. Boháč, M. Rott, and V. Kovář, "Text punctuation: An inter-annotator agreement study," in *International Conference on Text, Speech, and Dialogue*. Springer, 2017, pp. 120–128.

[177] T. Hastie and R. Tibshirani, *Generalized additive models*. Wiley Online Library, 1990.

[178] C. R. Mehta and N. R. Patel, "IBM SPSS exact tests," *SPSS Inc., Cambridge, MA*, 2010.

[179] M. E. Peters *et al.*, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[180] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Appendix A

Publications

International journals

[J1] Gy. Szaszák, M. G. Tulics, **M. Á. Tündik**, "Analyzing F0 discontinuity for speech prosody enhancement," *Acta Univ. Sapientiae Elect. Mech. Eng*, vol. 6, no. 1, pp. 59–67, 2014.

[J2] Gy. Szaszák, **M. Á. Tündik**, B. Gerazov, "Prosodic stress detection for fixed stress languages using formal atom decomposition and a statistical hidden Markov hybrid," *Speech Communication*, vol. 102, pp. 14–26, 2018.

[J3] A. Kovács, M. G. Tulics, **M. Á. Tündik**, A. Moró, A. Gróf, "Magmanet: Ensemble of 1d convolutional deep neural networks for speaker recognition in Hungarian," *Phonetician*, vol. 115, 2018.

[J4] **M. Á. Tündik**, V. Kaszás, Gy. Szaszák, "On the effects of automatic transcription and segmentation errors in Hungarian spoken language processing," *Periodica Polytechnica Electrical Engineering and Computer Science*, 2019.

Hungarian journals

[J5] A. Hilt, **M. Á. Tündik**, G. Bota, L. Nagy, K. Luukkanen, "Hívás közbeni beszédfordítás: új hangalapú szolgáltatás a telefonhálózatokban," *Híradástechnika*, vol. 72, pp. 10–15, 2017.

International conferences

[C1] Gy. Szaszák, **M. Á. Tündik**, K. Vicsi, "Automatic speech to text transformation of spontaneous job interviews on the HuComtech database," in *Proceedings of the 2nd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2011).* IEEE, 2011, pp. 1–4.

[C2] Gy. Szaszák, **M. Á. Tündik**, B. Gerazov, A. Gjoreski, "Combining atom decomposition of the F0 track and HMM-based phonological phrase modelling for robust stress detection in speech," in *Proceedings of the 18th International Conference on Speech And Computer (SPECOM2016)*, 2016, pp. 165–173.

[C3] **M. Á. Tündik**, B. Gerazov, A. Gjoreski, Gy. Szaszák, "Atom decomposition based stress detection and automatic phrasing of speech," in *Proceedings of the 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2016).* IEEE, 2016, pp. 25-30.

[C4] Gy. Szaszák, **M. Á. Tündik**, A. Beke, "Summarization of spontaneous speech using automatic speech recognition and a speech prosody based tokenizer," in *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016)*, 2016, pp. 221–227.

[C5] **M. Á. Tündik**, G. Kiss, D. Sztahó, Gy. Szaszák, "Assessment of pathological speech prosody based on automatic stress detection and phrasing approaches," in *Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2017).* IEEE, 2017, pp. 67–72.

[C6] **M. Á. Tündik**, B. Tarján, Gy. Szaszák, "A bilingual comparison of MaxEnt- and RNN-based punctuation restoration in speech transcripts," in *Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2017).* IEEE, 2017, pp. 121–126.

[C7] **M. Á. Tündik**, B. Tarján, Gy. Szaszák, "Low latency MaxEnt-and RNN-based word sequence models for punctuation restoration of closed caption data," in *International Conference on Statistical Language and Speech Processing.* Springer, 2017, pp. 155–166.

[C8] **M. Á. Tündik**, A. Hilt, G. Bota, L. Nagy, K. Luukkanen, "Access-independent cloud-based real-time translation service for voice calls in mobile networks," in *2018 11th International Symposium on Communication Systems, Networks Digital Signal Processing (CSNDSP)*, July 2018, pp. 1–6.

[C9] **M. Á. Tündik**, Gy. Szaszák, "Joint word-and character-level embedding CNN-RNN models for punctuation restoration," in *Proceedings of the 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2018)*. IEEE, 2018, pp. 135–140.

[C10] V. Kaszás, **M. Á. Tündik**, Gy. Szaszák, "A semantic space approach for automatic summarization of documents," in *Proceedings of the 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2018)*. IEEE, 2018, pp. 153–158.

[C11] **M. Á. Tündik**, Gy. Szaszák, G. Gosztolya, A. Beke, "User-centric evaluation of automatic punctuation in ASR closed captioning," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018)*. ISCA, 2018, pp. 2628–2632.

[C12] G. Járó, A. Hilt, L. Nagy, **M. Á. Tündik**, J. Varga, "Evolution towards Telco-Cloud: Reflections on Dimensioning, Availability and Operability," in *Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, 2019, pp. 1–8.

[C13] Gy. Szaszák, **M. Á. Tündik**, "Leveraging a character, word and prosody triplet for an ASR error robust and agglutination friendly punctuation approach," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*. ISCA, 2019, pp. 2988–2992.

[C14] **M. Á. Tündik**, V. Kaszás, Gy. Szaszák, "Assessing the semantic space bias caused by ASR error propagation and its effect on spoken document summarization," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*. ISCA, 2019, pp. 1333–1337.

Hungarian conferences

[C15] **M. Á. Tündik**, Gy. Szaszák, "Szövegalapú nyelvi elemzö kiértékelése gépi beszédfelismerö hibákkal terhelt kimenetén," in *Proceedings of the 12th Hungarian Conference on Computational Linguistics (MSZNY 2016)*, 2016, pp. 111–120.

[C16] **M. Á. Tündik**, G. Kiss, D. Sztahó, Gy. Szaszák, "Automatikus frázisdetektáló módszereken alapuló patológiás beszédelemzés magyar nyelven," in *Proceedings of the 13th Hungarian Conference on Computational Linguistics (MSZNY 2017)*, 2017, pp. 113–124.

[C17] **M. Á. Tündik**, B. Tarján, Gy. Szaszák, "Televíziós feliratok írásjeleinek visszaállítása rekurrens neurális hálózatokkal," in *Proceedings of the 14th Hungarian Conference on Computational Linguistics (MSZNY 2018)*, 2018, pp. 183–195.

[C18] **M. Á. Tündik**, Gy. Szaszák, "Kombinált központozási megoldások magyar nyelvre pehelysúlyú neurális hálózatokkal," in *Proceedings of the 15th Hungarian Conference on Computational Linguistics (MSZNY 2019)*, 2019, pp. 275–286.