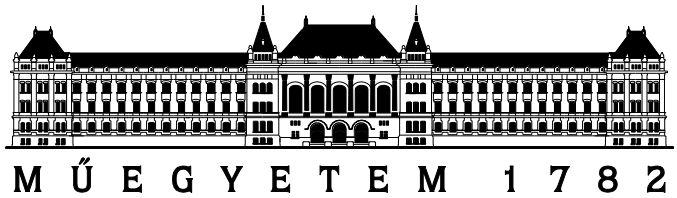

Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
Electrical Engineering Doctoral School

Recognition of Spontaneous Hungarian Speech without Language Specific Rules

Summary of the Original Contributions of the PhD Thesis

Péter Mihajlik, MSc

Research Supervisor:
Dr. Géza Gordos, DSc
Consultant:
Péter Tatai, MSc
Department of Telecommunications and Media Informatics

Budapest, Hungary
2010

1. Introduction

The research of automatic speech recognition (ASR) dates back several decades. The first speech recognition technique applicable for practical problems was the DTW (Dynamic Time Warping) method [Vintsjuk 68], [Myers & Rabiner 81], [Gordos & Takács, 83]. The principle is based on dynamic programming [Bellman 57]. DTW can be applied for language independent speaker dependent small vocabulary isolated word recognition. First the speech features are extracted and then this observation sequence is matched to templates. The closest match provides the recognition result. The main drawback of the technique is that templates must be taught by the user.

A significant progress was made by the introduction of HMM (Hidden Markov-Model) in ASR [Baker 75], [Jelinek & Bahl+ 75]. The basic processes of speech recognition did not change (feature extraction and pattern matching), however, the structure of background models became much more complicated and the model parameters were estimated off-line based on statistics. Speaker independent ASR could be approximated based on the appropriate statistics of hundreds or thousands of speakers' voices.

The next step in achieving LVCSR (Large Vocabulary Continuous Speech Recognition) was the integration of language models into the HMM framework [Jelinek & Mercer 80]. HMM state sequences can be used for acoustic modeling of the corresponding language or speech units (words, syllables or phones) and transition probabilities can be interpreted as conditional word transition probabilities. The recognition result of a generalized HMM recognition network is approximated typically by the best path of the network in the ML (Maximum Likelihood) sense. The best path can be calculated using the Viterbi-method [Ney 84].

The initial equation of continuous speech recognition is as follows:

$$\hat{W} = \arg\max_{W} P(W \mid O), \tag{1}$$

where $W = w_1,...,w_K$, $K \in N$ is a permitted word sequence and $O = o_1,...,o_T$ denotes the observation sequence of feature vectors extracted from the input speech. $\hat{W} = \hat{w}_1,...,\hat{w}_{\hat{K}}$, $\hat{K} \in N$ is the estimated (hypothesized) word sequence – the result of the recognition.

Applying the Bayes-rule on (1) we get the MAP (Maximum A Posteriori) equation of speech recognition:

$$\hat{W} = \arg\max_{W} P(W) \cdot P(O \mid W) \tag{2}$$

Equation (2) separates illustratively the language model $P(W)$ and the acoustic model $P(O \mid W)$. The low-level (coarticulation) and high-level (word-to-phoneme) pronunciation models are considered as parts of the acoustic model.

Thus, in order to develop a given language speech recognizer the acoustic and language models have to be generated. Although the generally used methods are largely based on statistics, language specific expert knowledge is required for the traditional creation of speech recognition models, too.

Each language has its own characterization, though, Hungarian is considered as a peculiar case. It has connections to the Finno-Ugric languages and to the Turkish languages as well – both branches are far from the Indo-European languages. Hungarian is known for its highly agglutinating nature and inflexion of verbs and nouns is also common. Theoretically, Hungarian has no diphthongs or triphtonghs; it applies vowel harmony within words and the language has several other special characteristics (e.g., there is no gender of nouns, etc.)

The recognition of spontaneous speech is a big challenge in itself due to the way of articulation and to other related phenomena, e.g., a-grammatical sentences, disfluencies, etc.

So, LVCSR of spontaneous Hungarian can be considered as a difficult and uninvestigated task. In addition, only a few publications are available dealing with continuous Hungarian speech recognition in general. Only [Szarvas 03] use context-dependent phone models and morpheme based language models for a dictation task. Other approaches apply context-independent phone models [Tóth 2009], [Szaszák 2008], [Bánhalmi & Paczolay+ 08], [Zsigri & Tóth+ 04], [Vicsi & Szaszák 04]. Most of the publications use word based language models. Considering other agglutinative languages (Finnish, Estonian, Turkish, Arabic) the difficulties due to high morphological complexity are successfully reduced by applying morpheme-like lexical units instead of words [Kurimo & Creutz+ 06], [Afify & Sarikaya+ 06]. However, to the best of my knowledge no positive improvement result is obtained for spontaneous speech; only one unsuccessful trial is reported for Egyptian Colloquial Arabic [Creutz & Hirsimäki+ 07]. Grapheme based speech recognition proved to be competitive for several languages [Kanthak & Ney 02], [Killer & Stüker+ 03], however, their integration with morpheme based modeling is typically ad-hoc and no detailed comparison with the classical word-phoneme approaches is available.

In the following the research objectives, the methodology and results related to the LVCSR of spontaneous Hungarian speech are introduced. The – somewhat unexpected – conclusion is that it is possible to achieve competitive results using an entirely data driven speech recognition approach. In other words, the application of language specific rules or the deeper understanding of the Hungarian language is not absolutely necessary to develop an efficient LVCSR system for spontaneous Hungarian.

2. Research Objectives

The general aim was the accurate and tractable automated recognition of Hungarian speech. The primary aim of this dissertation is finding answers for the Hungarian-related speech recognition modeling issues.

Issues addressed in the thesis:

I. Phonetic coarticulation modeling for Hungarian LVCSR. Is triphone modeling effective for Hungarian? Does Hungarian need special techniques for context dependency modeling?
II. Phonological coarticulation modeling of Hungarian. How to model pronunciation rules explicitly, such as assimilation, merging of consonants, voice harmonization, dropping and insertion of phonemes across words boundaries as well. Is it necessary to model them explicitly?
III. Lexical modeling of spontaneous Hungarian speech. How to cope with huge vocabularies, high OOV (Out Of Vocabulary) rate and data sparseness for language modeling. What kinds of subword lexical units are able to increase speech recognition accuracy and how to obtain them?
IV. Pronunciation modeling of spontaneous Hungarian speech. How can be the word-to-phoneme mapping automated. Do we really need several language specific rules and/or manual work for that?

No further investigation was planned beyond the above mentioned issues. The language modeling principle (N-gram approximation) and the low level acoustic modeling (via Gaussian Mixture Models) are considered as language independent components; therefore these are not discussed in the dissertation.

3. Methodology

The applied methodology of the dissertation follows the international standards. In the following, speech databases, recognition tasks, conditions and evaluations used are detailed.

3.1. Speech Databases and Tasks

The largest available Hungarian language speech databases were selected.

Telephone speech was used for the investigation of phonetic and phonological coarticulation modeling. The following Hungarian language telephone speech databases were merged in the experiments: MTBA [Vicsi & Tóth 02], Besztel, SpeechDat and Tesztel [Vicsi et al.]. These databases contain mainly read speech but spontaneous utterances can be found in them, too. Both fixed and mobile recordings have been made, altogether from 1600 speakers (about 50 hours of speech).

Spontaneous speech was used for the experiments that aimed at the investigation of lexical and pronunciation modeling. Only the MALACH (Multilingual Access to Large Spoken Archives) database [B1, J1] was appropriate for the purpose. It contains conversations with elderly Holocaust survivors. The topic was their personal stories from the period of World War II. The speech of the interviewees were sometimes incoherent, rich in disfluencies but some of the speakers spoke surprisingly well. The amount of transcribed speech material is 34 hours from 114 speakers. The sampling frequency was 48 kHz for the recorded speech.

Recognition tasks
Basically speaker independent large vocabulary continuous speech recognition tasks were defined since these problems are considered as the most challenging ones (apart from low signal-to-noise conditions). Some application oriented isolated word tests were also performed.

3.2. Training and Test Sets

The *training sets* contain the majority of the database recordings. 500-900-1300 speakers' data were used in the case of telephone speech and 104 speakers in the case of the spontaneous database. Four different size training sets were defined for telephone speech tasks: the smallest set is the manually segmented part of the MTBA database (~3 hours, 6000 recordings). The next one contains nearly the whole MTBA database (19 000 recordings). In the third training set most of the Besztel database is added to the second training set (39 000 recordings). Finally, the largest telephone training set is the extension of the previous set with the SpeechDat's 400 speakers data (44 000 recordings, 30 hours). In the case of the spontaneous task only one training set was defined with 26 hours of manually transcribed speech.

For the definition of *test sets* a fundamental requirement was that all test speakers must have been unseen for all of the training sets. 220 speakers, 2400 utterances were selected for telephone speech recognition experiments and 10 speakers, 8 hours of speech for the spontaneous tests. Furthermore, test sets were partitioned into well or weakly matching subsets. Considering telephone speech, the matched subset fits very well to the training sets in terms of word and diphone distribution whereas the unmatched set is completely mismatched. In the case of spontaneous MALACH database, the difference between test subsets is minor. In this thesis summary the differentiation of subsets is generally not denoted (in opposite to the dissertation).

3.3. Experimental Configurations

Experiments on telephone speech data: in each different coarticulation modeling approach the low-level acoustic models (triphones) have been rebuilt from scratch in all training sets. These acoustic models were used for continuous and isolated word tests and were evaluated on the weak and well matched test subsets.

Experiments on spontaneous speech: lexical modeling and pronunciation modeling experiments were run on the MALACH data with only one training set and with two test sets.

3.4. Experimental Conditions

The default settings of the speech recognition experiments are detailed as follows. (Any difference from the default settings will be emphasized later.)

Feature extraction: basic features were 12 MFCC (Mel Frequency Cepstral Coefficients) [Mermelstein 76] in the case of telephone speech, and 17 PLP (Perceptual Linear Prediction) [Hermansky 90] coefficients in the case of spontaneous speech. Dynamic Delta and Delta-Delta features were also calculated and added to the basic feature vector (delta windows size was +-2 frames and linear regression was applied). Static energy feature was suppressed in the telephone features and so 38 and 54 dimension final feature vector sizes were obtained. Blind channel equalization was applied in the telephone speech recognition experiments [Mauuary 98], [C15].

Simple cepstral mean subtraction was performed in the spontaneous speech recognition tests.

Atomic acoustic models: hidden Markov-model states with fixed transition probabilities were applied as lowest level acoustic models. Maximum number of Gaussian mixtures [Titterington & Smith+ 85] was 10 per HMM state. These models were trained based on the ML (Maximum Likelihood) principle applying EM (Expectation Maximization) algorithms [Dempster & Laird+ 77]. (Initialization, Viterbi training, embedded Baum-Welch reestimations, mixture splitting, etc. [Young 06].)

Phonetic coarticulation models: three state left-to-right HMM structure was used for both context-dependent and context-independent phone models. By default, context-dependent models were obtained by building ML phonetic decision trees for each state. About 100 phonetic categories were used as questions at the decision tree construction.

Context-dependency model: cross-word triphone models were applied.

Phonological coarticulation model: by default, *no explicit phonology model* was applied either in the training or in the tests. This is called the implicit model.

Phonological pronunciation models: In general, phoneme level pronunciations were derived from the orthographic word forms based on grapheme-phoneme rules [J6]. Pronunciation exceptions were manually phonemized in the spontaneous speech database if the phonetic form could not be derived by simple grapheme-phoneme rules. Pronunciation weights were estimated by the relative frequencies of pronunciation alternatives. Allophonic variants were not modeled. Long and short consonants were not differentiated. So, altogether 39 phonologic categories were used. The silence model was a three state context-independent model.

Vocabulary (lexical model): Default lexical units were words without pronunciation alternatives in the telephone speech task, and with weighted alternative pronunciations in the spontaneous MALACH task. The same vocabulary with a size of 1334 was used in the isolated word recognition tasks for both the well matched and the unmatched test subsets. Similarly, the same 5.6k vocabulary was used in both continuous telephone speech tasks. In the telephone speech recognition tests there were no out of vocabulary words. In the case of MALACH database the vocabulary size was 20 000 and OOV (Out Of Vocabulary) rates were around 15% for both test sets.

Language models: Standard N-gram language models were applied. 3-gram models were used in the telephone speech recognition tasks with Katz-back off [Katz 87] and Good-Turing [Good 53] smoothing. The language model training text was composed from the text of the well matched test set, but each type of sentence occurred only once in the training set. So, a word perplexity of 40 was measured on the well matched test set and 6230 on the unmatched set. Considering the MALACH database, only the transcription of the training set was used for language model training. The modified interpolated Kneser-Ney smoothing [Chen & Goodman 98] was applied on the spontaneous task. Optimal orders of N-grams were found 3 for word-based models

and 4 for morph based models. The word perplexity of the whole test set was measured as 336.

Recognition network construction: the integration of knowledge sources and the optimization of the resulting network were performed in the WFST (Weighted Finite State Transducer) framework [Mohri & Pereira+ 02]. We applied determinization on the phoneme-level network as optimization.

Decoding: All of the recognition experiments were run on the same 3GHz Pentium IV machine equipped with 2GB of RAM. The dynamic programming based decoding was performed with the VOXerver recognition engine. Pruning was set conservatively: recognition accuracies were highly saturated even if the RTF went relatively high in all experiments. In the telephone speech recognition tasks the pruning thresholds were fixed whereas in the spontaneous experiments the RTFs were kept at a constant level.

3.5. Evaluation of the Results

The basis of evaluation was distance measurement from manually transcribed test text. The following metrics were applied.

Metrics: The recognition result – the sequence of hypothesized recognition units (words, letters) – was compared to the reference by a dynamic programming technique where the following weights were defined:

C (correct recognition): 0
S (substitution): 10
D (deletion): 7
I (insertion): 7

The basis of the evaluation was the smallest overall weight. The most important recognition metrics were:

$$\text{Recognition accuracy } (\text{"}Acc\text{"}) = \frac{N - S - D - I}{N} \times 100\%, \qquad (3)$$

$$\text{Error rate } (\text{"}ER\text{"}) = \frac{S + D + I}{N} \times 100\% \qquad (4)$$

where *N* is the number of recognition units in the reference.

As the concrete recognition units were words and letters, the following metrics were applied in the experiments:

- WER (Word Error Rate): is the most widely accepted ASR evaluation metric.
- LER (Letter Error Rate): correlates better with the cost of manual corrections. In the case of morphologically rich languages it serves as a more precise metric than WER for measuring improvements. In the experiments whitespace was considered as a letter, too.

Practically, the relative improvement is the most illustrative indicator. In this study it is calculated both for WER and LER as follows.

$$\text{Relative Improvement } (-\Delta ER_{rel}) = \frac{ER_{reference} - ER_{new}}{ER_{reference}} \times 100\% \quad (5)$$

It can also be important to control the computation time. Real time factor RTF is defined as:

$$RTF = \frac{time\ length\ of\ computation}{length\ of\ recognized\ speech} \quad (6)$$

Significance tests: Statistical hypothesis tests were performed in order to control the significance of the improvement results. Wilcoxon signed rank tests [Kanji 94], [Daniel 78] – according to NIST's recommendation – were applied.

Based on the annotations, utterances or sequences of several utterances were considered as independent events. WERs and LERs calculated on these parts of the test databases were considered as independent random variables.

By evaluating the results the significance level of $p=0.05$ (confidence level of *0.95*) was chosen. Significant improvements are emphasized with *italic* type fonts.

4. Summary of the Original Contributions

4.1. Modeling Phonetic Coarticulations for Automatic Speech Recognition of Hungarian – Statement Group I.

Coarticulation – the interaction between consecutive speech sounds (or phones) – is a fundamental process of speech production (and perception). In *phonetic* coarticulations the phoneme identities of the interacting speech sounds are not theoretically affected. Statistical speech recognition does take this phenomenon into consideration, however, for Hungarian a rather implicit approach is applied as most widespread. Context-independent (monophone) models are used in the large majority of publications [Tóth 09], [Szaszák 08], [Bánhalmi & Paczolay+ 07], [Tóth 06], [Vicsi & Velkei+ 05]. Context-dependent phone models [Szarvas 03] and syllable-based models in [Czap 05] are also applied in but these merely introduce initial results. No study could be found by the author that compares context-dependent and independent approaches on known Hungarian speech databases.

One of the first research objectives was to clarify if Hungarian has any distinctive property (theoretic absence of diphthongs, for example) that explains the widespread use of context-independent phone models. Or whether phonetic context-dependent phone modeling can improve speech recognition accuracy significantly.

Context-dependent phone modeling has to face with the following issue. On the one hand it would be necessary to differentiate all phonetic context that result in different articulations of the given speech sound. On the other hand, model complexity has to be limited in order to avoid under-training of atomic acoustic models.

In the first statement the results and conclusions of a special context-dependent phone modeling technique are summarized. The essence of the approach – developed by our research group – is as follows. First linguistically motivated rules are used for clustering HMM states based on their phonological context. Then, if a resulting cluster has fewer training samples than a threshold, the degree of context dependency is reduced until enough training data can be assigned to the current cluster.

The technique is called a back-off triphone state clustering process since the handling of data sparseness is similar to the technique of back-off n-gram language model smoothing [Katz 87].

Statement I.1: [B2, B3, C7, C8] *I have experimentally shown that using the back-off triphone state clustering technique for automatic recognition of Hungarian speech can result in significant improvement of the recognition accuracy as compared to the context-independent phone modeling approach.*

The speech recognition results measured on a more than 1500 speakers database are shown in Table 1.

Table 1
Summary of context-independent and back-off context-dependent phone modeling results on telephone speech data

Test set	Average word recognition accuracy [%]		Average error rate reduction [%]
	Reference: context-independent phone modeling	Back-off triphone state clustering technique	
Isolated word recognition			
Matched	85.7	95.0	*65*
Unmatched	82.7	91.5	*51*
Continuous speech recognition			
Matched	80.3	90.8	*53*
Unmatched	20.5	41.6	*26*

As can be seen, the relatively simple back-off triphone modeling technique caused drastic improvements. Although only average results are shown, all improvements were significant even with the smallest (~3 hours) training set, as well as with the larger sets[1].

Despite the real time factors being different[2], the results can be compared to each other. As it has been already mentioned, recognition accuracy-RTF curves were well saturated in the experiments due to "conservative" pruning settings. In all tests the recognition speed was below real-time (RTF<1).

Next, two different context-dependent phone modeling approaches are compared with each other.

[1] Results are detailed in Section 4.4 of the Dissertation. See Tables 4.2, 4.3, 4.5 and 4.6.
[2] RTF measurement results are detailed in Tables 4.4 and 4.7 of the Dissertation.

Statement I.2: [B2]. *I have experimentally shown that significant improvement can be obtained in terms of recognition accuracy by using the principally data driven ML phonetic decision tree-based triphone state clustering method [Young & Odell+ 94] instead of the basically rule-based back-off clustering method.*

The corresponding recognition results can be found in Table 2. The aim of this thesis is to highlight the importance of data driven techniques in speech processing. Although both context-dependent modeling techniques use phonetic knowledge, the more data driven and widely used method performs significantly better in all train-test configurations than the other technique developed specifically for Hungarian. [3]

Table 2
Summary of context-dependent
phone modeling results on telephone speech data

Test set	Average word recognition accuracy [%]		Average error rate reduction [%]
	Reference: Back-off triphone state clustering technique	ML decision tree based triphone state clustering technique	
Isolated word recognition			
Matched	95.0	96.3	*26*
Unmatched	91.5	93.6	*24*
Continuous speech recognition			
Matched	90.8	92.5	*19*
Unmatched	41.6	50.0	*14*

Real time factors were different in the tests but in this series of experiments the better approach was also faster[4].

The most important conclusion of the first statement group is that the explicit modeling of phonetic coarticulation via context-dependent phone models is highly recommendable for Hungarian language, as well.

4.2. Modeling Phonological Coarticulations (Pronunciation Rules) for Automatic Speech Recognition of Hungarian – Statement Group II.

In phonological coarticulations at least one of the interacting speech sounds changes its phoneme identity. These phenomena are considered as language dependent and often called "pronunciation rules". On the one hand the need for modeling phonological coarticulations is apparently obvious since pronunciation modeling is typically based on phonemes. On the other hand, the deep differentiation between phonetic and phonological coarticulation has not proved to be essential in speech processing and its necessity may be questionable.

The explicit modeling of phonological coarticulations was found to be important in earlier studies. Pronunciation variants of consonants and vowels were suggested to be modeled by alternative allophonic realizations [Cohen 89]. Later, the application of

[3] See numerical recognition results in Tables 4.2, 4.3, 4.5 and 4.6 of the Dissertation.
[4] RTF measurement results are detailed in Tables 4.4 and 4.7 of the Dissertation.

phoneme level pronunciation alternatives became general. Pronunciation rules were formalized in generalized finite state machine frameworks and 4 – 8 % (relative) improvement have been achieved due to their application [Kaplan & Kay 94], [Mohri & Sproat 96], [Hazen & Hetherington+ 02].

However, as [Lamel & Adda 96] points out, too many pronunciation alternatives can make the recognition network highly confused, which can result in deteriorated recognition performance. Later [Jurafsky & Ward+ 01]'s work ("What kind of pronunciation variation is hard for triphones to model?") states that only syllable or higher level changes are hard triphones to model. Then research focused on implicit pronunciation modeling [Hain 02], [Kanthak & Ney 02], [Killer & Stüker+ 03], and the results discouraged rule-based pronunciation modeling (including explicit phonological coarticulations).

Considering Hungarian language, various studies investigate phonological coarticulations and related phenomena [Gósy 98] [Vicsi & Szaszák 04] [Zsigri & Tóth+ 04], [Tóth 09]. Recognition error rate reduction due to pronunciation modeling, however, is not typically reached. The most prominent exception is [Szarvas 03], where phonological coarticulations are modeled across word boundaries, as well, and a recognition accuracy improvement is achieved in a Hungarian language dictation task. The generalization of these results though is difficult because of the experimental conditions (small databases, no significance tests, etc.).

In the following sections I introduce my results achieved in the field of modeling Hungarian language phonological coarticulations for ASR purposes. The next statement is a positive example where explicit modeling of pronunciation rules does improve the results, whereas the second states that under more realistic conditions a simple implicit model can be competitive.

Statement II.1: [J2, B3, B4, C6, C9, C10] *I have experimentally shown, that – assuming a phonetic transcription of the training database where phonological coarticulations are taken into consideration (e.g., by manual phonetic transcription) – the explicit, across-word modeling of phonological coarticulation in the recognition tests can improve the recognition accuracy significantly over the baseline where the modeling of the phenomena is ignored.*

Table 3
Continuous speech recognition results on telephone speech data. Training data were phonetically transcribed manually.

Test set	Word recognition accuracy [%]		Error rate reduction [%]
	Reference: no phonology model in tests	Explicit phonological coarticulation modeling	
Matched	91.4	92.6	*14*
Unmatched	49.5	51.1	*3.2*

The experiments were run on the Hungarian language telephone speech database detailed in Section 3. The types of phonological coarticulation modeled were as follows:

P_1: Voicing assimilations /obligatory/
P_2: Merging + shortening of consonants /obligatory/
P_3: Partial assimilations controlled by the place of articulations /optional/
P_4: Full assimilations controlled by the place of articulations /optional/

The explicit phonology coarticulation model, P, is obtained by the following sequence of WFST composition operations:

$$P = P2 \text{ o } P4 \text{ o } P3 \text{ o } P2 \text{ o } P1 \quad (7)$$

This model is able to take cross-word phoneme-level coarticulatory effects into account as well.

The operation of the phonological transducer, P, can be illustrated by the following example.

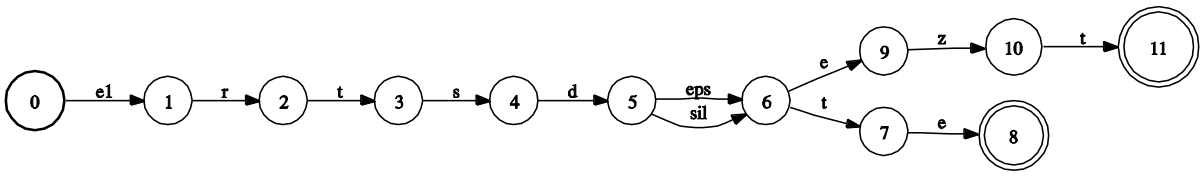

Figure 1. Connected word phoneme-level recognition network (F) ignoring phonological coarticulations. (Words are represented as phoneme-level finite state transducers: „*értsd te*" <you should understand it>, „értsd ezt" <understand this>.)

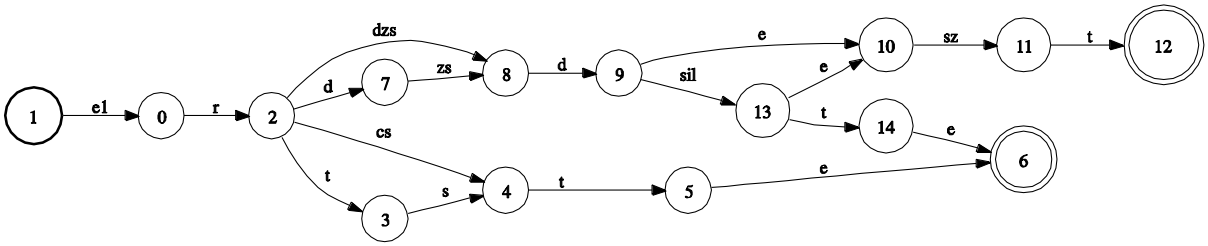

Figure 2. Connected word phoneme-level recognition network with explicit modeling of phonological coarticulations (P o F – the surface phonemic form of the automaton of Fig. 1.)

As Table 3 shows, P can successfully reduce phonological coarticulation mismatch between training and test conditions. In practice, however, it is not feasible to perform phonetic transcription manually. The next section summarizes my experiences with automated phonetic transcriptions.

Statement II.2: [C6], *I have experimentally shown that using implicit phonological coarticulation modeling – where training and test conditions are matched in terms of ignoring phoneme-level coarticulation rules – competitive recognition results can be obtained as compared to the case of (training-test) consequent explicit phonology modeling approach.*

Corollary: *The unweighted explicit modeling of typical Hungarian phonological coarticulations may be avoided without the significant loss of speech recognition performance, reducing significantly the complexity of speech recognition systems.*

Table 4
Summary of continuous speech recognition results on telephone speech data. Phonological modeling is automated and matched between training and tests.

Test set	Average word recognition accuracy [%]		Average error rate reduction [%]
	Reference: Explicit phonological coarticulation modeling	Implicit phonological coarticulation modeling	
Matched	92.5	92.5	0.2
Unmatched	50.0	50.8	-1.6

The experimental results summarized in Table 4 were obtained with 4 different training sets. No significant difference could be observed in the error rates in any experimental conditions.[5] The real time factors were higher with the explicit technique[6].

Comparing the results of the first and second group of statements, it can be concluded that the modeling of phonetic coarticulation through context-dependent phone modeling is indispensable, whereas the explicit modeling of phonological coarticulations has little benefit, if any. The results may have not only practical importance but also can support a theoretic assumption, that is, phonological coarticulations should be considered as special cases of phonetic coarticulations and do not form separate phenomena (so, the adjective "phonetic" would be superfluous).

4.3. Lexical Modeling for the Recognition of Spontaneous Hungarian Speech – Statement Group III.

It is self-evident that words are decomposed into phonemes in contemporary LVCSR systems. Less obvious questions are whether words need to be split into various subword lexical units and how to do this segmentation. These problems are posed primarily in the case of morphologically rich languages. In these languages there is a strong demand to reduce the data sparseness of language models, the dictionary sizes and the ratio of out of vocabulary words. Lexical modeling discusses these issues.

Formally, subword-based speech recognition can be defined as the generalization of the word-based principle:

$$\hat{M} = \arg\max_{M} P(M)P(O \mid M) \tag{8}$$

$$\hat{W} = f(\hat{M}) \tag{9}$$

where yet undefined symbols are M, the sequence of subword (orthographic) lexical units, and f, an operator that performs simple text concatenations and deletions on the recognized sequence of subword lexical units.

[5] For detailed recognition results, see Table 5.2 of the Dissertation.
[6] Real time factors are presented in Table 5.3 of the Dissertation.

In the most researched languages (like English, French, German, Spanish, etc.) the dominant approach is to apply words as lexical units. In case of morphologically rich languages – like Finnish, Estonian, Hungarian, Turkish, and Arabic – the agglutination and inflection of stems and suffixes can cause drastic vocabulary growth, high OOV rates and sparse data for language models. Applying morpheme-like lexical units (morphs[7]) instead of words seems a promising approach for these languages since the previous problems can be effectively treated. The reported recognition accuracies, however, did not always improve. The most significant improvement was obtained for Finnish LVCSR tasks [Hirsimäki & Creutz+ 06], in other languages and tasks smaller improvements were achieved [Arisoy & Can+ 09] and, in some cases, the error rate increased [Creutz & Hirsimäki+ 07]. To the best of our knowledge, positive results were published only for the recognition of planned or undefined [Afify & Sarikaya+ 06] speech style and the only result available for morph-based recognition of spontaneous speech is negative (for Egyptian Colloquial Arabic [Creutz & Hirsimäki+ 07]) as compared to the word baseline.

As for the Hungarian language, no public result is known to the author that could improve word-based speech recognition accuracy due to a more advanced lexical modeling technique. [Szarvas 03] enhances morph-based recognition by adding morpho-syntactic rules and [Vicsi & Velkei+ 05] applied grammatical morphs but word forms are not reconstructed.

The novelty of the third statement group is twofold. On one hand the first significant improvement results due to morph lexical modeling are introduced for Hungarian, on the other hand these are among the first positive results for spontaneous speech.

Statement III.1: [J1, B1, C1, C2, C3, C4, C5] *I have experimentally shown that the application of morpheme-like subword lexical units (morphs) instead of words can result in significant improvement of the recognition of spontaneous speech.*

The experiments supporting the thesis were performed on the Hungarian MALACH database. The results measured on the full test set can be found in Table 5.

The next statement is an enhancement of the previous one. It shows that no explicit morpho-syntactic knowledge is needed for the significant improvement to be achieved due to lexical modeling.

Statement III.2: [J1, B1, C1, C2, C5] *I have experimentally shown that using morph lexical units generated by an unsupervised statistical method [Creutz & Lagus 05b] significant improvement can be obtained over the word baseline in spontaneous speech recognition. Additionally, the recognition accuracy of this technique can be insignificantly lower than that of a grammatical or a combined (statistics + grammar used for the decomposition) morph based approach in a spontaneous recognition task.*

[7] In this study, all kinds of subword lexical units will be called as "morphs", though, typically, the morphs applied in the experiments resemble to the surface representations of morphemes.

Table 5
Summary of spontaneous LVCSR results on the MALACH database with various lexical models

Lexical model	Vocab. size	Recognition accuracy [%]		Relative error rate reduction [%]	
		Word	Letter	Word	Letter
Word – reference	20k	45.5	72.9	-	-
Stat. morph (MB)	4.6k	46.4	73.4	1.7	*1.8*
Stat. morph (MC-MAP)	5.5k	46.8	73.7	*2.4*	*3.0*
Gramm. morph (HSF)	8k	46.5	73.7	1.8	*3.0*
Gramm. morph (HCG)	6.7k	46.8	73.8	*2.4*	*3.3*
Combined morph (CHM)	6.7k	47.0	73.9	*2.8*	*3.7*

The following word-to-morph segmentation techniques were applied in the experiments:

1.) Unsupervised statistical morphs:

- ***MB (Morfessor Baseline)****:* The method aims at finding the optimal lexicon and segmentation, i.e., a set of morphs that is concise, and moreover gives a concise representation for the data [Creutz & Lagus 05a]. This model is inspired by the Minimum Description Length (MDL) principle [Risannen 78].
- ***MC-MAP (Morfessor Categories-MAP)****:* Aims at improving the segmentation obtained using the Baseline method. The morphs are tagged with category labels: prefix, stem, and suffix. By learning morph categories as well as sequential dependencies between these, the segmentation can be refined [Creutz & Lagus 05b]. The average length of the resulting morphs can be controlled by a perplexity threshold.

2.) Grammatical morphs – generated by using language dependent morpho-syntactic rules and databases [Trón & Németh+ 05], [Trón & Halácsy+ 06]:

- ***HSF (Hunmorph Strict Fallback)***: First the input word is analyzed as a non-compound word and strict morpho-syntactic rules must be satisfied. If no valid analysis is obtained in the first round, the word is assumed to be a compound word. Finally, if compound analysis fails, too, a heuristic guessing algorithm is applied.
- ***HCG (Hunmorph Compound Guessing)***: The input word is assumed as if it could be a compound word, and the heuristics mentioned previously are applied at the same time. Each input word is segmented in one pass. The number of alternative outputs can be much higher than in the case of the HSF method.

In case of multiple analyses the one containing the most morph segments is chosen.

3.) Combination of statistical and grammatical methods:

- ***CHM (Combined Hunmorph Morfessor):*** the MB algorithm is used to disambiguate the multiple morph analyses of Hunmorph system. The details of the technique can found in [C5]; it is used here as an upper reference.

As can be seen in Table 5, the unsupervised statistical MC-MAP, the grammatical HCG as well as the CHM method could significantly outperform the traditional word-based lexical modeling approach both in terms of word and letter error rates. No significant difference could be observed between the results of the previously mentioned three techniques. For precise comparisons, real time factors in all experimental conditions were kept in the interval of 4.2-4.3.

4.4. Pronunciation and Acoustic Modeling for the Recognition of Spontaneous Hungarian Speech – Statement Group IV.

In the classical speech recognition approach the orthographic words are first mapped to phoneme sequences then the phonemes to context-dependent phone segments. For these mappings typically several language specific rules and expert knowledge are used, e.g., grapheme-to-phoneme rules, pronunciation exception dictionaries and phonetic categories (nasals, velars, etc.). Statement II.2 showed that language specific rule-based modeling of phonological coarticulation was not effective in improving the recognition of Hungarian speech. In this section we go further: all language specific rules are excluded hoping no serious loss of the recognition accuracy.

The essence of the so-called "grapheme-based" speech recognition is that the phoneme-level is completely ignored – the words are segmented trivially to alphabetic letters (characters) and the acoustic models are built directly on these context-dependent graphemes. The same ML decision tree building technique can be used for graphemes as for phonemes. The phonetic categories used for decision tree building can be converted easily to graphemic categories. This approach was applied successfully for German and Spanish [Kanthak & Ney 02]. [Killer & Stüker+ 03] went even further by using "singleton" graphemic categories achieving fully data driven speech recognition with competitive results.

Grapheme-based acoustic and pronunciation modeling fits well to subword-based language modeling even if subword lexical units are obtained by using statistical algorithms. Recently, morph-based language modeling is applied increasingly with grapheme acoustic models, however, no earlier study is known for the author that compares the phoneme- and grapheme-based recognition with subword lexical units. Furthermore, considering Hungarian language speech recognition, no grapheme-based ASR result is published apart from [Zgank & Kacic+ 2005] where only initial command word results were introduced.

A novelty of statement group IV is that grapheme- and phoneme-based LVCSR of Hungarian is compared with each other and the investigations are made with subword lexical models[8]. The pronunciation modeling of the spontaneous task is indeed challenging: as Table 6 shows, the exception dictionary sizes and coverages are uncommonly large.

[8] In this extract of the Dissertation the result of only one type of subword lexical unit is introduced. For a comprehensive evaluation over acoustic and lexical units, please, refer to [J1] or to Tables 6.2 and 7.2 of the Dissertation.

Table 6
Exception dictionary sizes and coverage data of the MALACH Hungarian training speech database. Weighted exceptions dictionary is a part of the exception dictionary.

Lexical model	Full vocabulary size	Exception dictionary		Weighted exception dictionary	
		Size	Coverage [%]	Size	Coverage [%]
Word	20k	1743	47.1	720	46.2
Morph (MC–MAP)	5.5k	492	27.3	163	26.9

Statement IV. 1: [J1, C3] *I have experimentally shown that the recognition accuracy of context-dependent grapheme based speech recognition can be insignificantly lower than that of the classical phoneme based approach in Hungarian spontaneous LVCSR (if morph units are applied in the lexicon).*

Corollary: *The absence of manually made exception dictionaries and grapheme-to-phoneme rules in Hungarian language speech recognition systems does not necessarily result in significantly degraded recognition performance. Thus, these knowledge sources should not be considered as essential components for Hungarian speech recognition.*

The mapping of context-dependent graphemes to "physical" HMM states was performed by the same ML decision tree building technique as was already applied in statement I.2. The phonetic categories necessary for the decision tree building were converted to graphemic categories in the same way as described in [Kanthak & Ney 02]. Related results measured on the MALACH database are summarized in Table 7.

A question can be raised whether the only remaining language specific expert knowledge applied – phonetically based grapheme classes – can be omitted from a competitive speech recognition system for spontaneous Hungarian. The answer is in the next section.

Statement IV. 2: [J1] *I have experimentally shown that by using context-dependent grapheme-singleton acoustic models – where only trivial, one element graphemic classes are used in the ML decision tree building – it is possible to obtain insignificantly lower speech recognition accuracies in a spontaneous Hungarian LVCSR task than in the case of the classical phoneme-based approach (if morph units are applied in the lexicon).*

Corollary: *The recognition accuracy of Hungarian speech recognition does not necessarily degrade significantly if language specific rules and expert linguistic knowledge are not applied explicitly. Thus, language specific rules and expert linguistic knowledge in their abstract form (category definitions, rules, dictionaries, etc.) should not be considered as essential for Hungarian language speech recognition.*

Table 7
Spontaneous Hungarian LVCSR results with MC-MAP (statistical) morph lexical model and with various acoustic models on the MALACH database.

Acoustic model	Recognition accuracy[%]		Relative error rate reduction [%]	
	Word	Letter	Word	Letter
phoneme – reference	46.8	73.7	-	-
grapheme	46.2	73.5	-1.1	-0.7
grapheme singleton	46.3	73.6	-0.9	-0.3

For the results see Table 7. A singleton class – used for context dependency modeling – means that only one grapheme belongs to each class (an "s" or "y" grapheme, for example). In other words, there is no prior knowledge about the typical articulation of the speech sound corresponding to the given grapheme. All in all, no language-specific expert knowledge or rules are used to train grapheme-singleton acoustic models.

The summary of language-specific rules applied in the training of the previously discussed acoustic models is illustrated below.

- Phoneme-based model (reference):
  - ML weighted alternative pronunciations ("miért" <why>):
    miért 0.011 m é
    miért 0.426 m é r
    miért 0.269 m i é r
    miért 0.292 m i é r t
  - Foreign and traditional words' phonetic transcription:
    Churchill cs ö r cs i l
    Kossuth k o s ú t
  - Grapheme-to-phoneme conversion rules:
    cz c
    ch# cs
    ck# k
    ly j
    (# means word boundary symbol)
  - Phonetic classes:
    NASAL: m, n, ny
    FRONT: e, é, i, í, ö, ő, ü, ű

- Grapheme-based model:
  - (Gra)phonetic classes:
    NASAL: m, n, n, y
    FRONT: e, é, i, í, ö, ő, ü, ű

- Grapheme singleton based model:
  –
  (no language specific rule)

As Table 7 shows, the exclusion of simple and alternative pronunciation exceptions and grapheme-to-phoneme rules caused only insignificant error rate increase. Moreover, the removal of the remaining language specific expert knowledge (phonetic categories) even improved slightly the recognition accuracy. So, the entirely data driven approach performed only marginally and insignificantly worse in terms of word and letter error rates than the phoneme-based reference with the given subword lexical model.

4.5. Recognition of Spontaneous Hungarian Speech without Language Specific Rules – Statement Group V. (Synthesis)

Statement V. 1: [J1] *I have experimentally shown that in the case of spontaneous Hungarian speech recognition it is possible to achieve competitive results – as compared to the classical word-phoneme baseline*[9] *– without the explicit application of language specific expert knowledge*[10].

Table 8
Summary of classical (word-phoneme) and special, data driven (morph-grapheme) LVCSR results on the MALACH Hungarian database

Lexical-acoustical model	Recognition accuracy [%]		Relative error rate reduction [%]	
	Word	Letter	Word	Letter
Word – phoneme	45.5	72.9	-	-
Stat. morph (MC-MAP) – grapheme-singleton	46.3	73.6	1.5	*2.6*

As can be seen, the widespread classical approach could be outperformed both in terms of word and letter error rate by the fully data driven technique. Moreover, the letter error rate reduction of the language specific rule free approach was significant, as well. The improvement is possibly due to the morph-based lexical modeling method that takes the structure of Hungarian language speech into consideration implicitly.

The validity of statement group III, IV and V were confirmed – on higher recognition accuracy levels – by further researches applying speaker adaptations [J1, B1, C1, C2, C3, C4, C5].

[9] In the classical word-phoneme based ASR approach typically several language-specific rules and expert knowledge are used as illustrated in the previous subsection.
[10] The fully data driven approach applies morph lexical units generated by unsupervised learning, morph n-gram statistical language models and grapheme-singleton acoustic models which are described in earlier subsections.

The Application of the Results

The application of the results is nearly self-evident since either each result simplifies the construction speech recognition systems for Hungarian or improves the recognition accuracy.

The results of first statement group – the modeling of phonetic coarticulation of Hungarian – is expected to be used for significantly more precise recognition of Hungarian speech even with simple technologies.

The results of the second statement group allow a simplified construction of a continuous Hungarian speech recognizer because they show the inefficacy of the complex explicit modeling of phonological coarticulation.

The third statement group's results can be applied in spontaneous speech recognition for improved accuracy, primarily for Hungarian or for more complex languages. Besides, lower resources are required thanks to the smaller morph vocabulary.

Perhaps the most important application of the fourth statement group's results is that it encourages the development of speech recognition application for new languages. Thus, it showed for a highly complex language (i.e. Hungarian) and on a difficult task that language specific expert knowledge is not necessarily an essential component of a speech recognition system, and consequently, it can be omitted reducing significantly the development cost and time.

Finally, the statement group five, as a synthesis of the preceding results, can be applied for competitive, cost and time efficient speech recognition system development for appropriate Hungarian tasks. The achievement is due to data driven techniques and to taking the nature of Hungarian morphology and writing into account.

Most of the results of the PhD thesis are already used in industrial applications.

References

[Afify & Sarikaya+ 06] Afify, Mohamed; Sarikaya, Ruhi; Kuo, Hong-Kwang Jeff; Besacier, Laurent; Gao, Yuqing (2006): "On the use of morphological analysis for dialectal Arabic speech recognition", In INTERSPEECH-2006, pp. 1444-1447

[Arisoy & Can+ 09] Ebru Arisoy, Dogan Can, Siddika Parlak, Hasim Sak and Murat Saraclar. Turkish Broadcast News Transcription and Retrieval. IEEE Transactions on Audio, Speech, and Language Processing, 17(5):874-883, July 2009

[Bahl & Jelinek+ 83] L. R. Bahl, F. Jelinek, R. L. Mercer. A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, pp. 179–190, March 1983.

[Baker 75] J. K. Baker. Stochastic modeling for automatic speech understanding. In Reddy, R., editor, Speech recognition, pp. 512–542, New York, USA, Academic Press, 1975.

[Baum & Eagon 67] L. E. Baum, J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology. Amer. Math. Soc. Bull., Vol. 73, pp. 360–362, 1967.

[Bánhalmi & Kocsor+ 05] Banhalmi, A., Kocsor, A., Paczolay, D.: Magyar nyelvű diktáló rendszer támogatása újszerű nyelvi modellek segítségével, in Proc. of MSZNY 2005, pp. 337 – 347, Szeged, 2005.

[Bánhalmi & Paczolay+ 08] Banhalmi, A., Paczolay, D., Toth, L., Kocsor, A.: Investigating the robustness of a Hungarian medical dictation system under various conditions, International Journal of Speech Technology, VOLUME 9, ISSUE 3-4 (2008), PAGE 121-131.

[Bellegarda & Nahamoo 90] J. R. Bellegarda, D. Nahamoo. Tied mixture continous parameter modeling for speech recognition. IEEE Trans ASSP, Vol. 38, No. 12, pp. 2033–2045, December 1990.

[Bellman 57] R. E. Bellman. Dynamic Programming. Princeton University Press, Princeton, USA, 1957.

[Beulen & Ney 98] K. Beulen and H.Ney, Automatic Question Generation for Decision Tree Based State Tying, Proceedings of the ICASSP, pp- 805-808, Seattle, WA, 1998.

[Chen & Goodman 98] Stanley F. Chen and Joshua T. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[Cohen 89] M. H. Cohen. Phonological structures for speech recognition. Ph.D. dissertation, University of California, Berkeley, USA, 1989.

[Creutz & Lagus 05a] Creutz, M. and Lagus, K., "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.", Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March, (2005)

[Creutz & Lagus 05b] Creutz, M. and Lagus, K., "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text", In Proceedings of AKRR'05, Espoo, Finland, 15–17 June, (2005)

[Creutz & Hirsimäki+ 07] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, & A. Stolcke, Morph-based speech recognition and modeling of out-of-vocabulary words across languages, ACM Transactions on Speech and Language Processing 5(1), 2007.

[Czap 05] Czap L.: Audiovizuális beszédfelismerés és szintézis, PhD értekezés, BME, Budapet, 2005.

[Daniel 78] W. Daniel, Applied Nonparametric Statistics, Houghton Mifflin, 1978.

[Dempster & Laird+ 77] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal Royal Statistical Society, Series B, Vol. 39, No. 1, pp. 1–38, 1977.

[Good 53] Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika, 40(3 and 4):237-264.

[Gordos & Takács 83] Gordos G., Takács Gy. (1983) *Digitális beszédfeldolgozás*, Műszaki Könyvkiadó, Budapest.

[Gósy 98] Gósy Mária. A zöngésségi hasonulás a (spontán) beszédben. Beszédkutatás 1998, Ed. Gósy Mária, Akadémiai kiadó, Budapest, pp. 1-20, 1998

[Hain 02] T. Hain. Implicit pronunciation modeling in ASR. Proc. ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language, pp. 129–134, Estes Park, Colorado, USA, September 2002.

[Hazen & Hetherington+ 02] Timothy J. Hazen, I. Lee Hetherington, Han Shu and Karen Livescu, "Pronunciation modeling using a finite-state transducer representation," Proceedings of ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation, Estes Park, Colorado, September, 2002

[Hermansky 90] H. Hermansky. (1990) Perceptual linear predictive (PLP) analysis of speech, *Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738-1752.

[Jelinek & Bahl+ 75] F. Jelinek, F. Bahl, R. L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Trans. Information Theory, 21(3), pp. 250–256, 1975.

[Jurafsky & Ward+ 01] Jurafsky, Dan – Ward, Wayne – Jianping, Zhang – Herold, Keith – Xiuyang, Yu – Sen, Zhang. “What kind of pronunciation variation is hard for triphones to model?”, in IEEE ICASSP-01, Salt Lake City, Utah, 2001, pp. I.577–580.

[Kanji 94] G. Kanji, 100 Statistical Tests, SAGE Publications, 1994

[Kanthak & Ney 02] S. Kanthak, H. Ney. "Context-Dependent Acoustic Modeling Using Graphemes for Large Vocabulary Speech Recognition". In IEEE International Conference on Acoustics, Speech and Signal Processing, Vol 1, pp. 845-848, Orlando, FL, May 2002. download PostScript

[Kaplan & Kay 94] Kaplan, R. M. & Kay, M. (1994). 'Regular Models of Phonological Rule Systems'. Computational Linguistics 20, nr 3, 332-387.

[Katz 87] S. M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 35, No. 3, pp. 400–401, March 1987.

[Kirchoff & Vergyri+ 06] K. Kirchhoff, D. Vergyri, J. Bilmes, K. Duh, and A. Stolcke. 2006. Morphology-based language modeling for Arabic speech recognition. Computer Speech and Language, 20(4):589–608.

[Killer & Stüker+ 03] M. Killer, S. Stüker, and Tanja Schultz. Grapheme based Speech Recognition. Proc. Eurospeech, Geneva, Switzerland, September 2003

[Kwon & Park 03] O.-W. Kwon and J. Park. 2003. Korean large vocabulary continuous speech recognition with morpheme-based recognition units. Speech Communication, 39(3–4):287–300.

[Leggetter & Woodland 95] C.J. Leggetter and P.C. Woodland. (1995) "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression, " Proc. ARPA Spoken Language Systems Technology Workshop, Austin, TX, pp. 110-115.

[Levinson & Rabiner+ 83] S. E. Levinson, L. R. Rabiner, M. M. Sondhi. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Techn. Journal*, Vol. 62, No. 4, pp. 1035–1074, April 1983.

[López & Graña+ 03] López, K., Graña, M., Ezeiza, N., Hernández, M., Zulueta, E., Ezeiza, A. and Tovar, C., "Selection of Lexical Units for Continuous Speech Recognition of Basque", Proc. of CIARP, Havana, Cuba (2003) 244–250

[MacQueen 67] J. B. MacQueen. (1967) "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

[Mauuary 98] L. Mauuary. (1998) Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition, *Proc. EUSPICO'98*, Vol.1, pp. 359-363.

[Mermelstein 76] P. Mermelstein. (1976) Distance measures for speech recognition, psychological and instrumental, *Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed.*, pp. 374–388. Academic, New York.

[Mohri & Sproat 96] Mehryar Mohri and Richard Sproat. An Efficient Compiler for Weighted Rewrite Rules.In 34th Meeting of the Association for Computational Linguistics (ACL '96), Proceedings of the Conference, Santa Cruz, California. Santa Cruz, California, 1996.

[Mohri 97] Mehryar Mohri. Finite-State Transducers in Language and Speech Processing. Computational Linguistics, 23:2, 1997.

[Mohri & Pereira+ 02] Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language, 16(1):69-88, 2002.

[Myers & Rabiner, 81] C. S. Myers and L. R. Rabiner. (1981) A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389-1409, September

[Ney 84] H. Ney. The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 32, No. 2, pp. 263–271, April 1984.

[Ney & Mergel+ 87] H. Ney, D. Mergel, A. Noll, A. Paeseler. A Data-Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. I, pp. 833–836, Dallas, USA, April 1987.

[Risannen 78] Rissanen, J. (1978), 'Modeling By Shortest Data Description', Automatica, Vol. 14, pp 465-471

[Schillo & Fink+ 00] C. Schillo, G. A. Fink, and F. Kummert, Grapheme based speech recognition for large vocabularies, in Int. Conf on Spoken Language Processing, Beijing, China, Oct. 2000, pp. 129 132.

[Shafran & Hall 06] I. Shafran and K. Hall. 2006. Corrective models for speech recognition of inflected languages. In Proc. EMNLP, Sydney, Australia.

[Singh & Raj+ 99] Singh, R., Raj, B., Stern, R. M.: Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models. in Proc. Int. Conf. on Spoken Language Processing. Vol. 1 (1999) 117-120

[Stolcke 02] Stolcke, A., "SRILM – an extensible language modeling toolkit", In Proc. Intl. Conf. on Spoken Language Processing, Denver (2002) 901–904

[Steinbiss & Tran+ 94] V. Steinbiss, B.-H. Tran, H. Ney. Improvements in Beam Search. Proc. Int. Conf. on Spoken Language Processing, Vol. IV, pp. 2143–2146, Yokohama, Japan, September 1994.

[Szarvas & Furui 02] Mate Szarvas and Sadaoki Furui "Finite-state transducer based Hungarian LVCSR with explicit modeling of phonological changes" Proc. ICSLP2002, Denver, U.S.A., pp.1297-1300 (2002-9)

[Szarvas 03] Máté Szarvas. "Efficient large vocabulary continuous speech recognition using weighted finite-state transducers - The development of a Hungarian dictation system" Ph.D. dissertation, TITECH, Tokyo, Japan, 2003.

[Szaszák 2008] Szaszák György: Szupraszegmentális jellemzők szerepe és felhasználása a beszédfelismerésben, PhD disszertáció, BME, 2008.

[Titterington & Smith+ 85] Titterington, D., A. Smith, and U. Makov (1985) "Statistical Analysis of Finite Mixture Distributions," John Wiley & Sons.

[Tóth 06] Tóth, L.: Posterior-Based Speech Models and their Application to Hungarian Speech Recognition, Ph.D. Dissertation, University of Szeged, 2006.

[Tóth 09] Tóth, L.: Beszédfelismeresi kísérletek hangoskönyvekkel, Proc. MSZNY, pp. 206-216, 2009.

[Tóth & Kocsor+ 04] Tóth, L., Kocsor, A., Gosztolya, G.: Telephone Speech Recognition via the Combination of Knowledge Sources in a Segmental Speech Model, Acta Cybernetica, Vol. 16, No. 4, 2004.

[Trón & Németh+ 05] Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy. and Varga, D., “Hunmorph: open source word analysis”, In Proc. ACL 2005 Software Workshop, (2005) 77–85

[Trón & Halácsy+ 06] Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon (2006), Morphdb.hu: Hungarian lexical database and morphological grammar, In: Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA, pages 1670--1673.

[Vicsi & Vig 98] Vicsi, K. - Vig, A.: Az elsõ magyarnyelvû beszédadatbázis, Beszédkutatás '98, MTA Nyelvtudományi Intézete, Budapest 1998, pp. 163-177

[Vicsi & Velkei+ 05] Vicsi K. Velkei Sz., Szaszák Gy., Borostyán G., Teleki Cs., Tóth Sz. L., Gordos G..: Középszótáras, folyamatos beszédfelismerőrendszer fejlesztési tapasztalatai, Proc. of MSZNY 2005, pp. 348 – 360.

[Vicsi & Tóth 02] Vicsi K., Tóth L. Kocsor A., Gordos G. Csirik J. (2002): MTBA - Magyar nyelvű telefonbeszéd adatbázis. Hiradástechnika 2002/8. sz. pp. 35-39.

[Vicsi & Szaszák 04] Klára Vicsi, György Szaszák: Examination of Pronunciation Variation from Hand-Labelled Corpora. TSD 2004: 473-480. Text, Speech and Dialogue, 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004, Proceedings. Lecture Notes in Computer Science 3206 Springer 2004, ISBN 3-540-23049-1.

[Vicsi et al.] Vicsi Klára et al. http://alpha.tmit.bme.hu/speech/databases.php

[Vintsjuk 68] T. K. Vintsyuk, „Speech discrimination by dynamic programming”, Kibernetika, Vol. 4, pp. 81-88, Jan.-Feb. 1968

[Wilcoxon 45] Wilcoxon, F. "Individual Comparisons by Ranking Methods." Biometrics 1, 80-83, 1945.

[Young 06] S. J. Young. The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, Cambridge, England, December, 2006.

[Young & Odell+ 94] Young, S. – Odell, J. – Woodland, P. Tree-based state tying for high accuracy acoustic modelling. DARPA Human Language Technology Workshop, pages 307–312, March 1994.

[Zgank & Kacic+ 05] Zgank, A. - Kacic, Z. - Diehl F. - Juhar, J. - Lihan, S. - Vicsi, K. - Szaszák, Gy.: Graphemes as basic units for crosslingual speech recognition,, COST 278 Workshop, 2005

[Zsigri & Tóth+ 04] Zsigri, Gy., Toth, L., Kocsor, A. Sejtes, Gy.: Az automata és kézi szegmentálás ejtésvariációk okozta problémai, Proc. MSZNY 2004.

Publications Related to the PhD Thesis

Journal papers

[J1] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh, T. Fegyó: Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task, *IEEE Transactions on Audio Speech and Language Processing*, Volume 18, Issue 6, pp. 1588-1600, 2010.

[J2] P. Mihajlik, T. Révész, P. Tatai: Phonetic Transcription in Automatic Speech Recognition, *Acta Linguistica Hungarica*, Volume 49, Issues 3-4, pp. 407-425, 2003.

Chapters in edited books

[B1] P. Mihajlik, T. Fegyó, B. Németh, Z. Tüske, V. Trón: Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages: Hungarian ASR for the MALACH Project, In: V. Matousek, P. Mautner (ed.): *Text, Speech and Dialogue*, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 2007, Proceedings, Lecture Notes in Computer Science, Volume 4629/2007, pp. 342-350.

[B2] Mihajlik P., Fegyó T., Tatai P.: A New Method for the Construction of Context-dependent Phone Models for Automatic Speech Recognition *(in Hungarian)*, In: M. Gósy (ed.): *Speech Research 2006*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 2006. pp. 218-230.

[B3] P. Mihajlik, P. Tatai, G. Gordos: Automatic Phonetic Transcription and Its Application in Speech Recogniser Training: A case study for Hungarian. In: P. Divenyi, S. Greenberg, G. Meyer (ed.): *Dynamics of Speech Production and Perception*, IOS Press, Amsterdam, NATO Science Series; I., 374., Life and Behavioural Sciences, 2006. pp. 245-262.

[B4] Mihajlik P., Tatai P.: Automatic Phonetic Transcription for Hungarian Language Speech Recognition *(in Hungarian)*, In: M. Gósy (szerk.): *Speech Research 2001*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 2001. pp. 172-185.

Conference papers[11]

[C1] B. Tarján and P. Mihajlik: On Morph-based LVCSR Improvements, *Proc. SLTU 2010*, May 3-5, 2010, Penang, Malaysia, pp. 10-16.

[C2] P. Mihajlik, B. Tarján, Z. Tüske, T. Fegyó: Investigation of Morph-based Speech Recognition Improvements across Speech Genres, *Proc. Interspeech 2009*, Sep. 6-10, 2009, Brighton, United Kingdom, pp. 2687-2690.

[11] Each English language conference paper was peer reviewed. Hungarian conference papers were accepted based on an extended abstract.

[C3] P. Mihajlik, T. Fegyó, Z. Tüske, P. Ircing: A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian, *Proc. Interspeech 2007*, August 27-31, 2007, Antwerp, Belgium, pp. 1497-1500.

[C4] Tüske Z., Mihajlik P., Fegyó T.: Improved LVCSR of Spontaneous Hungarian by Speaker Adaptations in the MALACH project *(in Hungarian), V. Conference on Hungarian Computational Linguistics* 2007. December 6-7, Szeged, pp. 47-55.

[C5] Németh B., Mihajlik P., Tikk D., Trón V.: Combination of Statistical and Rule-based Morphological Analyzers for Speech Recognition *(in Hungarian), V. Conference on Hungarian Computational Linguistics* 2007. December 6-7, Szeged, pp. 95-105.

[C6] Mihajlik P.: Coarticulation models in Hungarian Language Speech Recognition *(in Hungarian), IV. Conference on Hungarian Computational Linguistics*, 2006. December 7-8, Szeged, pp. 231-242.

[C7] T. Fegyó, P. Mihajlik, M. Szarvas, P. Tatai, G. Tatai: Voxenter™ – Intelligent Voice Enabled Call Center for Hungarian, *Proc. Interspeech 2003*, Sep. 1-4, 2003, Geneva, Switzerland, pp. 1905-1908.

[C8] T. Fegyó, P. Mihajlik, P. Tatai: Comparative Study on Hungarian Acoustic Model Sets and Training Methods, *Proc. Interspeech 2003*, Sep. 1-4, 2003, Geneva, Switzerland, pp. 829-832.

[C9] P. Mihajlik, T. Fegyó, P. Tatai, G. Gordos: Pronunciation Modeling in Continuous Number Recognition, *Proc. ECMCS 2001*, Sep. 11-13, 2001, Budapest, Hungary, pp. 330-333.

[C10] T. Fegyó, P. Mihajlik, P. Tatai, G. Gordos, Pronunciation Modeling in Hungarian Number Recognition, *Proc. Interspeech 2001*, Sep. 3-7, 2001, Aalborg, Denmark, pp. 1465-1468.

Further Scientific Publications (Related to Automatic Speech Recognition)

Journal papers

[J3] Németh G., Olaszy G., Bartalis M., Zainkó Cs., Fék M., Mihajlik P.: Preparation of Speech Databases for more Efficient Support of Research and Development Objectives *(in Hungarian), Telecommunications*, Vol. LXIII, 2008/5, pp. 18-24.

[J4] Tüske Z, Mihajlik P, Tobler Z, Fegyó T, Tatai P.: Analysis and Optimization of Speech Endpoint Detection Methods for Automatic Speech Recognition *(in Hungarian), Telecommunications*, Vol. LXI, 2006/3, pp. 59-67.

[J5] Szarvas M., Fegyó T., Mihajlik P., Tatai P.: Results in the Hungarian Language Large Vocabulary Connected Word Speech Recognition *(in Hungarian), Telecommunications*, Vol. LVI, 2001/6, pp. 31-36.

[J6] Szarvas M., Fegyó T., Mihajlik P., Tatai P.: Automatic Recognition of Hungarian: Theory and Practice, *International Journal of Speech Technology*, Volume 3, Numbers 3-4, pp. 237-251, 2000.

Chapters in edited books

[B4] Németh G., Olaszy G., Bartalis M., Kiss G., Zainkó Cs., Mihajlik P., Haraszti Cs.: Automated Drug Information System for Aged and Visually Impaired Persons, In: Miesenberger K, Klaus J, Zagler W, Karshmer A (ed.): *Computers Helping People with Special Needs*, Lecture Notes in Computer Science, Volume 5105/2008, Springer Berlin / Heidelberg, 2008. pp. 238-241.

[B5] Tüske Z., Simon M., Mihajlik P., Fegyó T.: Recognition of Emotions based on Acoustic Speech Features *(in Hungarian)*, In: Gósy M. (ed.): *Speech Research 2007*. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 2007. pp. 151-161.

[B6] Fegyó T., Mihajlik P., Tatai P.: A Comparative Study on the Training Techniques of Phone Models used in Speech Recognition *(in Hungarian)*, In: Gósy M. (ed.): *Speech Research 2002*. Institute for Linguistics, Hungarian Academy of Sciences,, Budapest, 2002. pp. 185-196.

Conference papers

[C11] Tüske Z., Simon M., Mihajlik P., Gordos G.: Automatic Recognition of Speech Emotions *(in Hungarian), V. Conference on Hungarian Computational Linguistics*, 2007. December 6-7, Szeged, pp. 81-91.

[C12] Tarján B., Györki M., Mihajlik P., Gordos G.: Results in the Hungarian Language Confidence Estimation for Speech Recognition *(in Hungarian). IV. Conference on Hungarian Computational Linguistics*, 2006. December 7-8, Szeged, pp. 243-254.

[C13] Tüske Z., Mihajlik P., Tobler Z.: A Novel Speech Endpoint Detection Method Combined with Noise Estimation to Improve Speech Recognition Efficiency *(in Hungarian), III. Conference on Hungarian Computational Linguistics*, 2005. december 8-9, Szeged, pp. 371-382.

[C14] P. Mihajlik, Z. Tobler, Z. Tüske, G. Gordos: Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech, *Proc. Interspeech 2005*, Sep. 4-8, Lisboa, Portugal, pp. 2677-2680.

[C15] Z. Tüske, P. Mihajlik, Z. Tobler, T. Fegyó: Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum, *Proc. Interspeech 2005*, Sep. 4-8, Lisboa, Portugal, pp. 245-248.