



A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian

Péter Mihajlik¹, Tibor Fegyó^{1,2}, Zoltán Tüske¹, and Pavel Ircing³

¹Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary

²AITIA International, Budapest, Hungary

³University of West Bohemia, Plzen, Czech Republic

mihajlik@tmit.bme.hu, tfegy@aitia.ai, tuske@tmit.bme.hu, ircing@kky.zcu.cz

Abstract

A coupled acoustic- and language-modeling approach is presented for the recognition of spontaneous speech primarily in agglutinative languages. The effectiveness of the approach in large vocabulary spontaneous speech recognition is demonstrated on the Hungarian MALACH corpus. The derivation of morphs from word forms is based on a statistical morphological segmentation tool while the mapping of morphs into graphemes is obtained trivially by splitting each morph into individual letters. Using morphs instead of words in language modeling gives significant WER reductions in case of both phoneme- and grapheme-based acoustic modeling. The improvements are larger after speaker adaptation of the acoustic models. In conclusion, morpho-phonemic and the proposed morpho-graphemic ASR approaches yield the same best WERs, which are significantly lower than the word-based baselines but essentially without language dependent rules or pronunciation dictionaries in the latter case.

Index Terms: spontaneous speech recognition, morphology.

1. Introduction

Morphologically motivated language models [1–5] as well as grapheme-based acoustic models [5,6] have been successfully applied to various speech recognition tasks. The joint use of these technologies [5], which we call morpho-graphemic approach has been, however, not thoroughly investigated, especially for spontaneous speech recognition.

In this paper we focus on the recognition of spontaneous speech primarily in agglutinative languages. Typical examples of such languages are Finnish, Estonian, Turkish and Hungarian, but Basque, Korean, Inuktitut, Swahili, and many other languages show similar features as well. The majority of them share two main characteristics. First, the number of word forms can be very high due to various stem and suffix agglutinations and inflections. Second, canonical pronunciations can typically be inferred from orthography using simple grapheme to phoneme conversion rules.

The obvious language modeling problem of these morphologically rich languages, i.e., the high number of rare word forms, is generally alleviated by using morphemes instead of words as basic units. An important question is how to segment the words into morphemes. A rule-based approach is used for newspaper reading in Hungarian [1], however, a direct comparison to word-based results is lacking. In [2] morpheme-like lexical units are obtained by rule-based tools, but they perform in Basque ASR worse than word-based ones. [3] successfully applies morphs resulted from statistics

and rule-based analysis for Korean LVCSR. Considerable error reductions are reported in [4] using a statistics based unsupervised morph segmentation tool called “Morfessor Baseline” for Finnish read newspaper speech recognition. Additionally, good results are achieved using the same statistical segmentation tool for Estonian and Turkish read speech recognition [5]. The latter may be the first occurrence of the morpho-graphemic approach in speech recognition.

In the following we present our work and results related to the large vocabulary spontaneous Hungarian MALACH speech recognition task.

2. Corpus and baseline system

2.1. The MALACH Hungarian speech corpus

The MALACH project (Multilingual Access to Large Spoken Archives) aims at providing improved access to the archived testimonies of Holocaust survivors. The testimonies were given in 32 languages and a considerable subpart (more than 2,000 hours) consists of Hungarian interviews. However, only 31 hours of this speech were transcribed so far and from now on we will refer to this specific part as the MALACH Hungarian speech corpus.

2.1.1. Speech and transcription data

Speech was recorded with a sampling frequency of 44.1 kHz in common environments (usually survivors’ homes). Given the nature of the interviews, MALACH speakers are typically aged and their speech is sometimes incoherent, rich in disfluencies and occasionally strongly accented. However, some of the interviewees speak close to the standard way.

During the transcription process, orthographic and phonemic variants were noted in parallel if pronunciation could not be automatically derived correctly from orthography. Phonological co-articulations were not considered during transcription.

2.1.2. Training and test sets

For training, 15-minute segments from 104 speakers were transcribed, starting from 30th minute of each interview (yielding a total of 26 hours, 200K running words). For test, a 5 hour (34K words, 210K letters) set was defined with variable length of transcribed data from 10 other speakers.

The test set was partitioned into several subsets. Matched subset for speaker independent recognition is defined as the set of test utterances collected after the 15th minute of the testimonies (about half of test recordings, 17K words, 108K

letters). Weakly-matched subset containing many named entities is defined as the complement of the matched subset, i.e., utterances from the first 15 minutes.

Test and adaptation subsets are also defined for speaker dependent ASR. Testimonies from one male and one female speaker were divided into an adaptation set comprising the first 15 minutes of each interview and a test set comprising the remaining hour (M/F: 5.5K/4K words, 35K/24K letters).

2.2. Baseline ASR system

A modified HTK front-end was used to get PLP-based acoustic features [7]. Speaker independent decision-tree state clustered cross-word triphone models with approximately 3000 HMM states were trained using ML estimation [8]. Three state left-to-right HMMs were applied with 10 mixture components per state.

Word to phoneme mapping was performed using simple rules for the majority of training text words [9]. Weighted alternative pronunciations were used only for those orthographic words which occurred in the training transcriptions with more than one different explicitly annotated pronunciation as exceptions.

Only training data transcripts were used for language modeling. Word-level 3-gram language model with modified, interpolated Kneser-Ney smoothing was built using the SRILM toolkit [10]. No pruning was applied in the language modeling or during the offline recognition network construction and optimizations performed primarily with the AT&T FSM toolkit [11]. One-pass decoding was done by an enhanced version of the decoder mentioned in [12] with a Real-Time Factor (RTF) of 2.8 on a 3GHz CPU.

Table 1. *Speaker independent baseline results [%]*.

Technique	Voc-ab.	Weak match		Matched	
		WER	LER	WER	LER
Baseline	20k	56.2	28.7	53.0	25.6

Letter Error Rates (LER's) were calculated as well, because in case of morphologically rich languages they sometimes can be more reliably used for evaluations than Word Error Rates (WER's). The OOV rate was about 15% for both test sets.

The results are comparable to the MALACH ASR systems for other languages even though the Hungarian training data are four times smaller than the other databases.

3. Morph-based ASR approaches

Word-based language modeling in morphologically rich languages seems somehow not well-grounded. E.g., a normal Hungarian word may correspond to three or even more English words because Hungarian glues cohesive morphs into one word while English writes them separately. Hence, morph-based ASR in case of agglutinative languages can be quite similar to the word-based one in English.

3.1. Using morphs as basic units of language

Basic sub-word lexical units will be called as morphs in the following regardless of their origin or meaning. Explicit word break symbols required for the restoration of word forms in the recognized string [4] are also considered as morphs.

Formally, morph-based ASR can be expressed as a straightforward generalization of word based- ASR:

$$\hat{M} = \arg \max_M P(M)P(O|M) \quad (1)$$

$$\hat{W} = f(\hat{M}) \quad (2)$$

where W means word sequence, M means morph sequence, O denotes acoustic feature vector sequence and f means trivial textual concatenation and deletion operations on the (possibly tagged) hypothesized morph sequence.

3.2. Morph segmentation techniques

The basic problem is how to determine the morph segmentation, i.e., f^{-1} which converts W to M . The following context insensitive word to morph segmentation techniques have been evaluated on transcribed MALACH Hungarian text.

3.2.1. Ocamorph – a linguistically motivated approach

Ocamorph is an open-source linguistic knowledge-based morphological analyzer with recently added morph segmentation capabilities [13]. The tool itself is not language dependent; however, the morph database and the associated rules currently exist only for Hungarian.

Two operating modes have been investigated for morph segmentations.

- **Strict Fallback (OSF):** first the input word is analyzed as a non-compound word and strict morpho-syntactic rules must be fulfilled. If no valid analysis is available in the first run, the word is assumed to be a compound word. Finally, if compound analysis fails, too, a heuristic guessing algorithm is applied.
- **Compound-Guessing (OCG):** the input word is considered as if it could be a compound word and also, heuristics are applied at the same time.

3.2.2. Morfessor families – statistical approaches

Morfessor tools use unsupervised data-driven methods that discover the regularities behind word formation in natural languages [14], [15].

The following two statistical morph segmentation tools have been applied in the experiments:

- **Morfessor Baseline (MB):** 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. This model is inspired by the Minimum Description Length (MDL) principle [14].
- **Morfessor Categories ML (MC-ML):** aims at improving the segmentation obtained using the Baseline method. The morphs are tagged with category labels and there are three categories in use: prefix, stem, and suffix. By learning morph categories as well as sequential dependencies between these, the segmentation can be refined [15].

Though the tools are able to take word frequencies into account, we use only the word forms as input information, i.e., each input word count is adjusted to 1.

3.2.3. Usage of segmentation tools

All segmentation tools are applied to the Hungarian MALACH manually transcribed training text as follows.

1. All text tokens from the preprocessed (lower cased, etc.) training text are collected into a list.
2. Spellings, foreign words and any non-word tokens are removed.

3. Morph segmentation is applied on the filtered list resulting in a word to morph ($W \rightarrow M$) dictionary.
4. Word break symbols (#) are inserted into the training text.
5. Words found in the $W \rightarrow M$ dictionary are replaced by (white space separated) sequences of morphs.

In this way word-based training text is transformed into a sequence of morphs in a broader sense where (possibly agglutinated) foreign words, word break symbols, spellings, hesitations and tagged morphs etc. are treated equally as simple morphs.

3.3. Morph-based language modeling

Once a word to morph mapping (f^{-1}) is available for the training texts, $P(M)$, i. e., the morph-based language model can be built. This can be done exactly in the same way as in case of words.

Unlike in [4] and [5] where morph N-gram order of 4-6 were found as best for book and newspaper readings recognition we found that a 3-gram morph LM performs best in terms of WER and LER on the spontaneous MALACH database. (See 3-gram morph LM results in Table 2, 3 and 4.)

3.4. Phoneme-based acoustic modeling of morphs

Similar to the word acoustic models, the morph acoustic model, $P(O|M)$, theoretically can be decomposed into morph to phoneme mappings (pronunciation model) and phoneme acoustic models (3). This we call morpho-phonemic approach:

$$\hat{M} = \arg \max_M P(M)P(\Phi | M)P(O | \Phi) \quad (3)$$

where Φ denotes sequence of phonemes.

The only requirement of this decomposition is the availability of $P(\Phi|M)$, i.e., the morph to phoneme mappings. The derivation of phoneme level pronunciations from orthographical morphs can be, however, problematic even in case of “phonetic writing” which is a typical property of rich morphology languages.

3.4.1. Morphs to phonemes mapping issues

First, trivially, hand made morph pronunciation lexicons can hardly be used alone, if morphs resulted from statistical segmentations. Hence, automatic methods are indispensable.

Second, even if grapheme to phoneme rules do exist for a given language their application to morphs can be problematic. Namely, in many languages (like German, Italian, Polish, Hungarian, etc.) typically a cluster of graphemes identifies one or more phonemes and if such a cluster is split phonemic pronunciations of the resulting morphs cannot always be determined correctly. Statistical segmentation methods, however, are blind to these grapheme clusters.

Furthermore, as it is well known, significant deletions can occur in the spontaneous pronunciations of common words. As [16] points out syllabic deletions can not be well modeled implicitly in the traditional triphone approach, hence explicit weighted alternative pronunciations should be associated to the given orthographic word. However, automatically non derivable pronunciations can be lost due to morph segmentations.

The situation is similar or worse if foreign and other exceptional words (abbreviations, etc.) are split. In these

cases even standard phonemic pronunciations can be lost due to splitting.

3.4.2. A phoneme-based approximation

We used a phoneme-based approximation to obtain morph to phoneme mappings. First, the word exception pronunciation dictionary is simply applied to the morphs of the training text. Then remaining morphs which were not found in the word exception dictionary are mapped to phonemes using simple grapheme to phoneme rules [9].

Though the method allows incorrect pronunciations, as well, practically it enables the usage of morph-based language models.

As can be seen on Table 2 all of the morph-based techniques outperform the word-based baseline system on the matched test set but only the improved statistical method (Morfessor Categories ML) gives better WER results on the weakly matched set. RTF was about 4 in each case.

Table 2. *Speaker independent morpho-phonemic ASR results [%].*

Technique	Voc-ab.	Weak match		Matched	
		WER	LER	WER	LER
OSF	8.0k	56.3	28.2	51.8	24.7
OCG	6.7k	56.7	28.3	51.7	24.6
MB	4.6k	56.2	28.5	51.3	24.7
MC-ML	5.5k	55.9	28.2	51.1	24.5

3.5. A morpho-graphemic approach

The application of phoneme-based acoustic models requires considerable amount of language specific knowledge (grapheme to phoneme rules, and/or manual phonemic transcriptions) and, “in return”, their integration with morph-based language models is difficult. Using graphemes instead of phonemes in acoustic models, however, causes no serious ASR performance degradation in “phonological writing” languages like German, Dutch and Italian [6]. At the same time grapheme acoustic models fit well to morph-based language modeling even if morph segmentation is obtained by statistical algorithms.

Formally, morpho-graphemic speech recognition approach can be similarly defined to morpho-phonemic one:

$$\hat{M} = \arg \max_M P(M)B(\Gamma | M)P(O | \Gamma) \quad (4)$$

where Γ denotes sequence of graphemes and B refers to binary “probability” distribution.

In our approach even foreign, traditional, etc. morphs grapheme “pronunciations” are obtained as their linear sequence of (alphabetic) letters, thus no alternative pronunciations are used.

In the experiments context-dependent grapheme acoustic models were trained similarly as phoneme acoustic models. Phonemic questions used in decision tree constructions for context dependent state tying were simply converted to graphemic questions as in [6].

Table 3. *Speaker independent word- and morpho-graphemic ASR results [%].*

Technique	Voc-ab.	Weak match		Matched	
		WER	LER	WER	LER
Word-Gr.	20k	57.1	28.7	53.9	25.8
MC-ML-Gr.	5.5k	56.9	28.3	51.7	24.6

Table 3 shows speech recognition results obtained with grapheme acoustic models coupled with traditional word based and with previously best performing morph-based language models. Though recognition errors are a bit higher than in the phonemic case, morpho-graphemic results are in average better than the baseline without using any manual or rule-based pronunciations.

4. The effect of acoustic model adaptations

We made speaker-wise tests to investigate the effect of acoustic model adaptations on various ASR approaches. For speaker adaptations (SA) global, supervised, ML linear mean and covariance transformations [8] were applied on the female (F) and male (M) speakers' data mentioned in 2.1.

Morph segmentation was performed with the Morfessor Categories ML algorithm [15]. In other aspects the ASR techniques were as described earlier. (Speaker independent reference results are denoted with "SI" on Table 4.)

Table 4. *Absolute word and letter error rates and their relative improvements - ΔW_{rel} and ΔL_{rel} , respectively - due to morph-based language modeling [%].*

F speaker – SI	WER	LER	ΔW_{rel}	ΔL_{rel}
Word-phonemic	51.9	21.7		
Morpho-phonemic	47.1	19.3	9.2	11.3
Word-graphemic	52.3	21.8		
Morpho-graphemic	48.6	20.0	7.2	8.0
F speaker – SA	WER	LER	ΔW_{rel}	ΔL_{rel}
Word-phonemic	47.2	18.3	16.4	19.4
Morpho-phonemic	39.4	14.8		
Word-graphemic	47.0	17.8	15.4	17.7
Morpho-graphemic	39.9	14.6		
M speaker – SI	WER	LER	ΔW_{rel}	ΔL_{rel}
Word-phonemic	49.1	23.7		
Morpho-phonemic	48.1	23.3	2.1	2.0
Word-graphemic	49.3	23.8		
Morpho-graphemic	48.8	23.5	0.9	1.3
M speaker – SA	WER	LER	ΔW_{rel}	ΔL_{rel}
Word-phonemic	46.3	21.2	4.2	5.6
Morpho-phonemic	44.4	20.1		
Word-graphemic	47.0	21.5	4.8	5.5
Morpho-graphemic	44.8	20.3		

As Table 4 shows speaker adaptation of acoustic models significantly increases the improvements due to morph-based language modeling and at the same time it practically eliminates the differences between best grapheme and phoneme based results. Moreover, it boosts the recognition process resulting in an RTF=2–3 for word and morph based techniques.

5. Conclusions

Various acoustic and language modeling approaches have been applied to a difficult Hungarian spontaneous speech recognition task. Best word and letter error rates were obtained using morph language models resulting from statistical segmentations and with speaker adapted acoustic models. The proposed morpho-graphemic approach significantly outperformed the traditional word-based system and gave essentially the same best recognition results as the morpho-phonemic one but without any hand made pronunciations or language specific morphosyntactic or grapheme to phoneme rules.

6. Acknowledgements

We thank for the support of grants NSF No. IIS-0122466, JHU 8202–48279, MSMT LC536, and NKFP-2/034/2004.

7. References

- [1] Szarvas, M. and Furui, S., "Finite-State Transducer based Modeling of Morphosyntax with Applications to Hungarian LVCSR", Proceedings of ICASSP, 368-371, Hong Kong, China, May, 2003.
- [2] 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, 244-250, Havana, Cuba, Nov, 2003.
- [3] Kwon, O.-W. and Park, J., "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," Speech Communication, Vol. 39, Issue 3-4, 287-300, Feb. 2003.
- [4] Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Pykkönen, J., and Virpioja, S., "Unlimited vocabulary speech recognition with morph language models applied to Finnish.", Computer, Speech and Language, Vol. 20, Issue 4, 515-541, October, 2006,
- [5] Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pykkönen, J., Aluma, T. and Saraclar, M., "Unlimited vocabulary speech recognition for agglutinative languages", HLT-NAACL, New York, USA, June 5-7, 2006.
- [6] Kanthak, S. and Ney, H., "Context-Dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition", In Proc. of ICASSP, 845-848, May, 2002.
- [7] Psutka, J., Ircing, P., Psutka, J. V., Radova, V., Byrne, W., Hajic, J., Mirovsky, J., and Gustman, S., "Large vocabulary ASR for spontaneous Czech in the MALACH project", EUROSPEECH-2003, 1821-1824
- [8] Young, S., Ollason, D., Valtchev, V., and Woodland, P., The HTK book (for HTK version 3.2.), March, 2002.
- [9] Szarvas M., Fegyó, T., Mihajlik, P., and Tatai, P., "Automatic Recognition of Hungarian: Theory and Practice", International Journal of Speech Technology, 3:277-287, December, 2000.
- [10] Stolcke, A., "SRILM – an extensible language modeling toolkit", In Proc. Intl. Conf. on Spoken Language Processing, pages 901–904, Denver, 2002.
- [11] Mohri, M., Pereira, F., and Riley, M., "Weighted Finite-State Transducers in Speech Recognition", Computer Speech and Language, 16(1):69-88, 2002.
- [12] Fegyó, T., Mihajlik, P., Szarvas, M., Tatai, P., and G., Tatai, "VOXenter - Intelligent voice enabled call center for Hungarian", In EUROSPEECH-2003, 1905-1908.
- [13] Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy. and Varga, D., "Hunmorph: open source word analysis", In Proc. ACL 2005 Software Workshop, 77-85
- [14] Creutz, M. and Lagus, K., "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.", Publ. in Comp. and Inf. Sci., Report A81, HUT, March, 2005.
- [15] Creutz, M. and Lagus, K., "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text", In Proc. of AKRR'05, Espoo, Finland, 15-17 June, 2005
- [16] Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., Sen, Z., "What kind of pronunciation variation is hard for triphones to model?", In Proceedings of ICASSP. Vol. 1. 577–580, 2001.