Acta Cybernetica **19** (2010) 695–713.

Speech Recognition Experiments with Audiobooks*

László Tóth[†], Balázs Tarján[‡], Gellért Sárosi[‡], and Péter Mihajlik[‡§]

Abstract

Under real-life conditions several factors may be present that make the automatic recognition of speech difficult. The most obvious examples are background noise, peculiarities of the speaker's voice, sloppy articulation and strong emotional load. These all pose difficult problems for robust speech recognition, but it is not exactly clear how much each contributes to the difficulty of the task. In this paper we examine the abilities of our best recognition technologies under near-ideal conditions. The optimal conditions will be simulated by working with the sound material of an audiobook, in which most of the disturbing factors mentioned above are absent. Firstly pure phone recognition experiments will be performed, where neural net-based technologies will also be tried as well as the conventional Hidden Markov Models. Then we move on to large vocabulary recognition, where morph-based language models are applied to improve the performance of the standard word-based technology. The tests clearly justify our assertion that audiobooks pose a much easier recognition task than real-life databases. In both types of tasks we report the lowest error rates we have achieved so far in Hungarian continuous speech recognition.

Keywords: speech recognition, LVCSR, audiobooks

1 Introduction

Creating speech recognizers that operate reliably in practice requires the handling of several factors that may not arise under laboratory conditions, but are inevitable in real life. The pressure from industry towards creating robust recognizers forces researchers to focus more on issues like the handling of noisy and spontaneous speech. This is of course understandable, but from a scientific point of view it is unusual to move on to a more difficult task before solving the simple one. And even

*This research was partially supported by TÁMOP-4.2.2/08/1/2008-0008, TÁMOP-4.2.2-08/1/KMR-2008-0007, OM-00102/2007 and by KMOP-1.1.3-08/A-2009-0006.

[†] Research Group on Artificial Intelligence, Hungarian Academy of Sciences and University of Szeged, E-mail: `tothl@inf.u-szeged.hu`

[‡] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, E-mail: {`tarjanb, sarosi, mihajlik`}`@tmit.bme.hu`

[§] THINKTech Research Center, Nonprofit LLC, Hungary

the recognition of clean and well-articulated speech has not yet been fully solved, and in particular for Hungarian we do not know how our recognizers would behave under such conditions. We think that the study of simpler recognition problems should not be abandoned: though the result of these are less applicable directly, they can shed light on how the various factors contribute to the difficulty of the recognition of real-life speech, and on how to handle these. Also, there might be applications where good quality speech can be assumed (for example, in a broadcast news captioning system both studio quality recording and good articulation are expected).

In this paper we test our current speech recognition systems on an audiobook. Our aim is to see how they perform under nearly ideal conditions. In Section 2 we explain why we regard the audiobook as providing 'ideal' speech by giving examples of the most important factors that are not present in the material of the audiobook. Then in Section 4 we present pure phone recognition experiments, which are useful for comparing the performance of various acoustic modeling techniques. In real life, however, one would expect a word-level output from a recognizer, so in Section 5 we create language models for the recognition task using both word- and morph-based techniques. In the last section we summarize our findings and draw some pertinent conclusions.

2 Factors Hindering Speech Recognition

In real-life situations several factors arise that degrade the performance of automatic speech recognizers. In this section we list the most important factors and try to assess their impact on the recognition rates in a real-life situation and also in the case of an audiobook.

It is hard to imagine such a real-life recognition environment where background noises could be perfectly excluded. One would find noises even in such high-quality recordings as broadcast news – e.g. paper being rustled and people taking breaths. And in fact there are quite a lot of applications that require speech recognition under a high level of noise, as in a car or cockpit. Many studies have been conducted to compare human and machine speech recognition under noisy conditions, and the results are usually disappointing [10]. The distorting effect of transfer media (most typically a phone line) is also counted as noise, and speech recognizers can be surprisingly sensitive to even a change in the microphone they are used with. Compared to the average background noise conditions and the quality and variability of telephone microphones and transfer lines, audiobooks contain practically no background noise and distortion, as they are normally recorded in sound-proof studios using professional equipment.

Speech recognizers can perform quite differently for different persons; that is, they are sensitive to the articulation characteristics and peculiarities of the speaker. We can demonstrate this by creating a histogram of the recognition accuracies for various speakers. The MTBA Hungarian Telephone Speech Corpus [23] is really suitable for such a test, as it contains recordings made from 500 people. Fig. 1

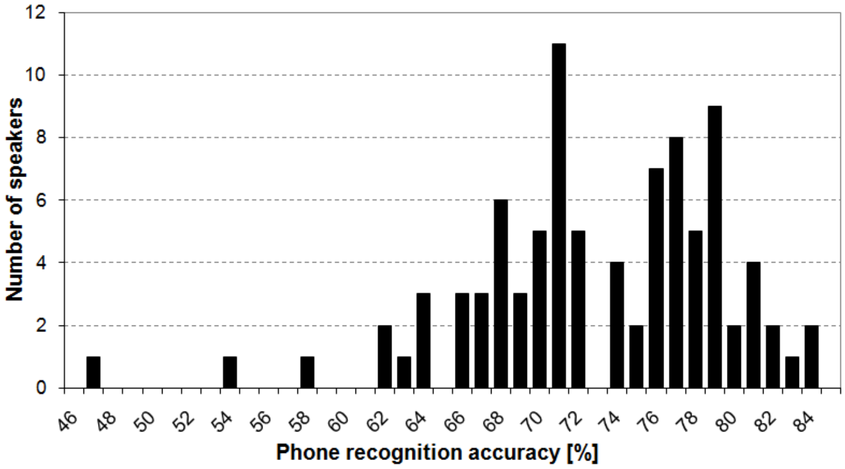

Figure 1: The distribution of phone recognition accuracy as a function of the speaker on the MTBA corpus

shows a histogram of phone recognition results obtained using this database that were presented in an earlier paper by the authors [21]. As can be seen, compared to the average value of around 74%, the results have quite a huge deviance in both directions. Though the recording conditions (phone line and background noise) also vary from speaker to speaker, we observed that the recordings are usually clean, so we think that the huge deviation in the scores can mostly be blamed on the sensitivity of the recognizer to the actual speaker's voice. Compared to the MTBA database, the sound material of the audiobook we used was presented by just one person, so the problems that could be caused by changing speakers were totally excluded.

The speech databases used to train speech recognizers were, for a long time, created by asking people to read aloud some written text. It was realized not long ago that there are huge differences between spontaneous articulation and careful pronunciation (e.g. reading aloud). Research has turned only recently towards studying real-life spontaneous speech. To assess the influence of spontaneous pronunciation on speech recognition, tests were made where people were asked to re-read their own passages that had been recorded at a meeting. The speech recognition error jumped by a factor of about two between the two recording (pronunciation) conditions. As regards Hungarian, Tarján et al. ran recognition tests on both planned and spontaneous speech using the same technology [20]. Although the results are not precisely comparable, as there were differences other than speaking style between the two recording situations, the error rate increase of a factor of two that they obtained accords well with the international findings. This drop in performance is so big that there are various applications where a trained speaker is applied to repeat the spontaneous parts – for example, in a Japanese system that uses speech recognition to close-caption broadcast TV news [27]. These problems cannot arise with an audiobook that is presented by an actor or actress. We think this is a guarantee for the cleanest possible articulation.

A further factor that may adversely affect the pronunciation is the presence of an emotional load. Similar to spontaneous articulation, the study of emotional speech has become popular only relatively recently. In Hungarian these studies have just been started [22], so we cannot give concrete examples of its influence on the recognition rates. In contrast with the above three factors, this one may be present to some degree in an audiobook, as the reading actor might use an emotional coloring of his voice to increase the expressive power of his speech. However, in the concrete recording we chose, such situations did not occur.

3 The Audiobook and its Preparation

Audiobooks have gained increasing popularity in the last couple of years. A lot of novels have been released in this ‘talking book’ format in Hungarian as well. We chose an audiobook for which the original novel is old enough so that its text is no longer copyrighted. Our choice fell on the short story collection by Gyula Krúdy entitled ‘Sinbad’s Voyages’, presented by the actor Sándor Gáspár. The total duration of the audiobook was 212 minutes, and we converted it to Microsoft Wave mono 16 bit PCM format with a sample rate of 16kHz. The whole book was carefully listened to, looking for differences between the original text and the sound material. Only a minimal amount of such differences were found, consisting mostly of the skipping or the insertion of short interjections such as ‘ah’. The music signals occurring at the end of each chapter were of course removed, and each file was segmented further into roughly two-minute long parts. Apart from these, no other modification was required.

For training and test purposes the recordings were divided into two parts. From the ten short stories eight were used for training (186 minutes) and two for testing (26 minutes).

4 Phone Recognition Experiments

In the first set of experiments the goal was pure phone recognition. Here no higher (word or morpheme-level) language model is applied, so the aim of recognition was to output a phonetic transcript that was as close to the signal’s real phonetic content as possible. This set of experiments was motivated by our recent phone recognition results on the MTBA database [21]. That is, we intended to apply the same methodology as was used there, and by comparing the results we hoped to gain a good insight into how much the factors mentioned above hinder the recognition performance. To aid the comparison, the results obtained with the MTBA corpus will be repeated here (a detailed description of the corpus can be found in [23]). The acoustic modeling technique used was the same as that we apply and discuss here.

In addition to the conventional hidden Markov modeling scheme, in the acoustic modeling step we applied a neural net-based technology as well. Hence, after presenting the data processing steps, we first give a theoretical description of this approach, and move on to the details of the implementation only afterwards. We round off this section with the comparison of the phone recognition scores obtained using the various techniques.

4.1 Preprocessing of the Data

In the feature extraction step, the most conventional 39-dimensional mel-frequency cepstral (MFCC) feature set (including the delta and acceleration coefficients) was extracted from the recordings [8]. The MTBA database was represented by PLP features in the reference that served as our comparison, because we had to comply with an English system in that paper [21]. However, previous experience tells us that the recognition results obtained with the two representations rarely show any significant difference.

Training and testing a phonetic recognizer requires a phone-level transcription of both the train and the test data. However, for the audiobooks only the orthographic transcription was readily available, which we had to convert to its most probable phonetic manifestation. This is usually done by collecting the different word forms occurring in the corpus, and transcribing them one by one. The word-level transcript given for each sentence is then mapped to the most probable sequence of phonetic segments by so-called forced alignment [8]. We performed this alignment using a set of phone models trained on the MRBA corpus [24] in earlier experiments.

Though Hungarian writing is almost phonetic, there are several difficulties which are not trivial to handle in an automatic phonetic transcription system. The first one is the case of the two-character letters, which can be identified only by morphologic analysis (a classic example is the word *pácsó*). The second one is that certain consonant clusters may or may not undergo assimilation depending on the position of morpheme boundaries. Lastly, in some cases assimilation is optional, and more than one pronunciations might be correct. A typical example of this are the word boundaries, where the consecutive words may be pronounced with a short silence between, but coarticulation and even assimilation may also occur.

Due to the above, a precise phonetic transcriber is not easy to create. Hence we decided to construct the phonetic transcript at the level of syllables instead of words. This idea was motivated by the fact that with careful pronunciation the vowels are not discarded or reduced (apart from shortening and lengthening). Second, assimilation affects only the consonant clusters and does not spread across vowels. Thus instead of creating a pronunciation dictionary of words, we created it for consonant clusters (for simplicity we will refer to these as syllables, though they are different from the linguistic definition of the syllable). We should add that the space character was also treated as a special consonant, so our ‘syllables’ were allowed to spread across word boundaries. The text of the audiobook contained 7186 unique word forms, but only 809 different syllables. These units were then transcribed manually. This transcription contained all the possible pronunciation alternatives of the syllable, so for example the letter-sequence 'T SZ' has the pronunciations [t sil s], [t s] and [t͡s], where *sil* denotes silence. The drawback of this method is that it does not see the whole word during transcription, and so it offers pronunciation alternatives that are not correct in the given context (for example, [paːt͡ʃoː] for the word *pácsó*). We hoped that in such cases the recognizer would be able to automatically choose the correct transcript during the forced alignment phase because of its better acoustic match. The phonetic label set used by the transcriber contained 52 symbols.

Besides the transcriber described above, an alternative phone-level transcription was also created using the full (language model-supported) recognizer configuration that will be presented in the next section. This resulted in a quite different series of symbols, because for those experiments a different phonetic transcriber was used. This was designed based on our earlier experimental findings that in large vocabulary recognition the explicit handling of the assimilations and similar phonological phenomena during the phonetic transcription of the dictionary is not necessary, provided that context-dependent phone models are applied at the phone level [14]. The explanation is that these models are implicitly aware of the phonetic context, and hence are sufficient to cope with most coarticulation artifacts. Therefore the phonetic transcriber used in these experiments performs only the most trivial grapheme-phoneme conversions, *and does not model such phonological processes as assimilation.* Also, the short and long consonants were not handled as separate labels by this converter, because fusing them causes only a negligible amount of confusions at the word level. Thus it operates with a label set of only 38 phonetic symbols.

4.2 Artificial Neural Net-Based Acoustic Modeling

The conventional Hidden Markov Modeling (HMM) technique approximates the probability distribution of the building blocks (model states) by fitting Gaussian mixtures (GMM) on the training data. In monophone modeling the states correspond to phone-thirds, while in the case of context-dependent triphone models the phonetic context is also reflected in the labeling of the building units (often referred to as senons [8]). This techniques raises the number of states from about one hundred to several thousands, but also brings considerable improvements in the performance.

Rather than refining the recognition units, an alternative approach is to improve the estimation of the probability distributions. One such solution is to use artificial neural nets (ANN) instead of Gaussian mixtures. The default training algorithm of neural nets is discriminative, and thanks to this it generally achieves slightly higher classification accuracies than the generatively trained GMMs. Secondly, the ANN-based systems are trained on several neighboring frames instead of just one, which again yields significantly better results. We should mention, however, that both discriminative training and the use of a longer observation context is possible with GMMs as well – it is just not part of the mainstream technology.

Under proper circumstances the output values of the neural net can be interpreted as probability estimates, and can be integrated into the conventional HMM scheme with some slight modifications to it. The resulting construct is known as the HMM/ANN hybrid model [2]. For smaller subtasks the hybrid was reported to outperform conventional HMMs. For example, the best phone recognition results on the TIMIT database were all obtained by applying neural nets [17, 15]. In larger systems, however, their applicability is much less obvious. First, their extension to context-dependent phone models is problematic, while on large (hundred hours) databases context-dependent modeling brings about a huge improvement for classic HMM models. Second, their combination with the standard language models seems to be less effective than that with conventional HMMs.

Most of these problems can be circumvented by using the so-called HMM/ANN tandem technology [6]. Here the neural net outputs are not interpreted as probability estimates, but as a non-linear transformation of the feature set. Hence, they can be used as the input for training a conventional HMM. With this trick only the acoustic preprocessor gets replaced and no other modification of the standard HMM recognizer is required. The only obvious drawback is that now there will be two training steps instead of one, and the evaluation will obviously also be slower.

Fig. 2 shows a schematic diagram of the processing steps involved for the conventional, the hybrid and the tandem methods.

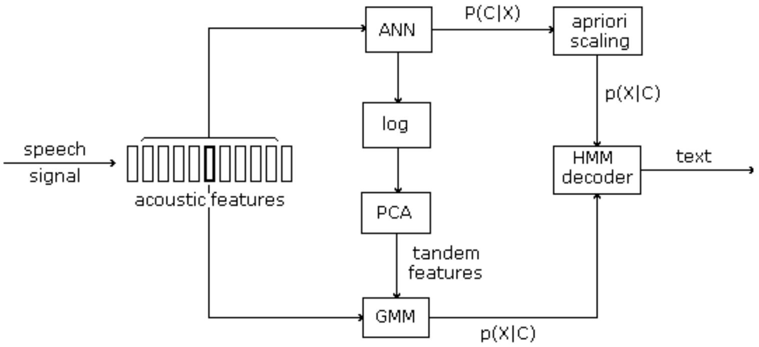

Figure 2: Schematics of the processing steps for the conventional HMM (lower path), the hybrid (upper path) and the tandem (middle path) models

4.3 Implementation Details

After the theoretical issues here we present the technical details of our implementation. The neural net we applied was a conventional feed-forward multilayer perceptron net with one hidden layer of 500 neurons. The number of outputs was the same as the number of phones, that is, 52 in the case of the syllable-based transcriber and 38 for the word-based transcriber. The output layer applied the softmax nonlinearity, while the hidden neurons utilized the sigmoid function. The number of inputs was 351, corresponding to 9 neighboring frames of MFCC vectors. The net was trained with backpropagation, and cross-validation on 10% of the training data served as the stopping criterion. The training targets were the phone labels obtained via forced alignment during the data preparation step. Before using the net outputs as features in the tandem, they were logarithmized and decorrelated using PCA. This eases the fitting of the data by Gaussian curves. A further trick we applied was to concatenate the tandem feature set with the MFCC feature set, and use them together. Though theoretically the two sets are highly redundant, the method still brings about a slight improvement in performance in practice.

As the implementation of the conventional HMM recognizer the well-known HTK package was used [26]. The phone models applied were standard 3-state left-to-right models with 9 Gaussians per state. In the case of the conventional HMM configuration both monophone and triphone models were trained. For the hybrid and tandem systems only monophone models were used. This is because efficient hybrid/tandem triphone modeling would require training the neural net with the senons as targets instead of the phones, and handling such a huge net is problematic. Extending ANN-based modeling to context-dependent cases is thus the most important topic of this area nowadays [1].

After the conventional maximum likelihood training the HMM models were further refined by the application of discriminative training, using the maximum mutual information (MMI) training criteria. Fortunately, the HTK package contains this algorithm.

As we mentioned previously, our goal here is pure phone recognition, so word-level (and morph-level) language models will be applied only in the next section. However, it is usual to support the recognition process at this low level by the application of phone bigrams. These were created from the force-aligned phonetic transcript of the training database.

4.4 Results and Discussion

Table 1 summarizes the phone accuracy scores obtained with the various acoustic modeling techniques, both for the syllable-based and the word-based automatic phonetic transcripts (the MTBA corpus was transcribed manually). For comparison the performance of the same models on the MTBA database is also shown, where the corresponding results were available (cf. [21]).

Comparing the results obtained with the syllable-based and the word-based transcripts, the former are consistently better for all but the simplest configuration. The accuracy obtained with the triphone model is, however, among the three highest, in spite of the fact that it was trained on the word-level transcript. This is in accordance with our earlier finding that context-dependent triphone modeling is able to handle not only the phonetic, but the phonological coarticulation effects as well, so it does not require the explicit modeling of the latter during phonetic transcription [14]. All the other models were trained with context-independent target labels, and they all show a degraded performance on the word-level transcripts.

Table 1: Phone recognition accuracies with the various acoustic models (the three highest scores are shown in boldface).

	MTBA (manual)	audiobook (syllabic)	audiobook (word-level)
HMM (monophone, no lang. model)	53.37%	72.18%	72.95%
HMM (monophone, phone bigram)	—	80.64%	76.85%
HMM (triphone, phone bigram)	—	—	**85.88%**
Tandem (monophone, no lang. model)	65.09%	79.49%	78.15%
Tandem (monophone, phone bigram)	69.67%	83.62%	80.93%
Tandem (mono., phone bg., discr. tr.)	73.93%	**86.26%**	82.84%
Hybrid (no language model)	—	84.84%	82.10%
Hybrid (phone bigram)	—	**86.60%**	82.69%

This indicates that monophone models do require the help of the phonetic transcriber. A further observation is that the phone-bigram consistently adds smaller improvements to the scores of the models trained with the word-level transcripts. A reasonable explanation is that the assimilation rules applied by the syllabic transcriber decrease the perplexity of the phone-bigram (consider, for example, the case of voiced-unvoiced consonant connections). This points towards the preference of full phonetic transcriptions when phone recognition output is required (but, of course, this argument has no importance when higher level language models are present).

As regards comparing the various modeling techniques, we can see that the tandem and hybrid models are indeed capable of the same (or even slightly better) performance than the conventional triphones, in spite of the fact that they use only monophone labels. Therefore, it would be very important to find such methods that are able to combine the advantages of the ANN-based and triphone technologies.

Now, comparing the audiobook results with those obtained with the MTBA corpus, one can find huge differences in each row of the table. These differences reflect how much easier it is to recognize the content of an audiobook than recordings made in real-life situations. The best result obtained on the MTBA is just slightly better than the worst score with the audiobook. We should also mention here that the situation could be even worse, because the MTBA corpus contains read speech (through phone lines). The recognition results of spontaneous phone calls would presumably be even worse. At the other end, the 86% accuracy score obtained on the audiobook is quite good: we made a test where a human subject was asked to read the output of the phone recognizer, and he reported that he was able to understand the whole story (one of the stories from the test set), apart from a couple of sentences. But, of course, an objective evaluation would require creating a language model, so that the whole recognition process would be automated. This is just what we shall do in the next section.

5 Large Vocabulary Speech Recognition Experiments

Phone recognition tests can be informative when we compare various acoustic models, but in a practical situation we of course expect a word-level output from a recognizer. In this section large vocabulary recognition test will be presented. On the acoustic modeling side only the conventional triphone HMM technique will be pursued, but in addition to the speaker-specific model training, comparisons will be made with models that have been trained on a larger, multi-speaker corpus. An adaptation of these models to the single speaker of the audiobook will also be attempted. On the language modeling side, we first present experiments with conventional word-level n-grams. In Hungarian the number of word forms is much higher than that in English, so the decomposition of the words into morphs can be beneficial. We have recently shown the usefulness of this technique on other domains [13], and we shall apply the same technology here. For the construction of a word- or a morph-based language model a large corpus of training text corpora is required. The various methods we applied for this and the parameters of the corpora assembled will also be presented. After, we will present our findings and discuss them.

5.1 Speaker-Independent, Speaker-Adapted and Speaker-Specific Acoustic Modeling

First a speaker independent state clustered cross-word triphone model was trained using ML (Maximum Likelihood) estimation [26]. Three-state left-to-right HMMs (Hidden Markov Models) were applied using GMMs (Gaussian Mixture Models) for each state. The model was trained on the MRBA database [24] augmented with 10 hours of transcribed press conference speech [20]. The feature type was MFCC (Mel Frequency Cepstral Coefficients) with delta and delta-delta parameters, which were calculated using blind channel equalization [11] at 8 kHz bandwidth. The resulting model contained 2100 HMM states with 7 Gaussians for each state. This speaker independent (SI) acoustic model will be referred to as **HMM (triphone, SI)**.

In the next step, speaker adaptation was applied using the unsupervised MLLR [9] technique. Speaker independent ASR was performed on the test and training set altogether, and the results were used as transcripts for a formally supervised adaptation. Thus the model parameters were retrained while the number of HMM states and the number of Gaussians per state were not changed. The speaker adapted (SA) acoustic model will be denoted by **HMM (triphone, SA)**.

In the third configuration the speaker independent model was used only for the phone labeling of the training set using forced alignment. Then an entirely new, speaker dependent (SD) model was trained from the MFCC features of the training data using the forced labels. This model had 1400 HMM states, each consisting a mixture of 10 Gaussians, and will be referred to as **HMM (triphone, SD)**. The same HMM-based triphone model was also used in Section 4 (third line of Table 1),

Table 2: Statistics of the language model training databases

	Size	Word Vocab.	Morph Vocab.	Word PPL	OOV Rate [%]
NM	17.6K	6.3K	2K	604	30.3
AM	1.4M	152K	18K	1792	2.9
SM	11.5M	590K	49k	2136	2.4

so we can make a fair comparison between the phone-level and the LVCSR (Large Vocabulary Continuous Speech Recognition) results.

5.2 Language Modeling

5.2.1 Collection and Preparation of the Text Corpus

The training transcript of the recorded literary novel contained fewer than 18K words. This transcription formed the smallest, **Novel Matched (NM)** corpus. Thereafter a separate training set – independent of the novel – was collected from the same author, Gyula Krúdy, which resulted in a 1.4M word size database. The source of this **Author Matched (AM)** corpus was the freely accessible Hungarian Electronic Library [28]. Finally, the corpus size was increased nearly ten times from the works of other authors from the early twentieth century, thus an 11.5M word count **Style Matched (SM)** corpus was created. This expansion includes texts from the Hungarian Electronic Library, the Digital Academy of Literature [29] and the Electronic Archive of Periodicals [30].

Following the text collection, the database had to be processed further. No extraordinary corpus preparation was required for word-based speech recognition, just the removal of any non-word characters, some number-to-text conversion and finally the conversion of all letters to lowercase. In contrast, the morph-based system required a special treatment of the given training text data. In our approach, first word boundary symbols <w> are placed into the text after each word, and are considered as separate morphs. (<w> symbols are required for the reconstruction of word boundaries in the decoder output [7]). Then segmentation is performed on the word dictionary by using the Morfessor Baseline (MB) algorithm [4]. MB is an unsupervised, language independent method for splitting words into morpheme-like lexical units called morphs. The method aims at the determination of the optimal lexicon and segmentation, that is, a set of morphs that is concise, and moreover gives a concise representation for the data. The corpus for a morph-based speech recognition system is obtained by replacing each word of the corpus by the corresponding morph sequence. For the statistical details of the training databases see Table 2 (PPL stands for perplexity and OOV denotes the out-of-vocabulary words).

5.2.2 Word- and Morph-Based Language Models

All the word- and morph-based n-gram language models were built on the corresponding database with *full vocabularies* applying the modified, interpolated Kneser-Ney smoothing technique [3] implemented via the SRILM toolkit [18]. Performance tests were run with several order of n-gram models – 2- to 4-grams – to find the optimal language model parameters. Based on these, full 3-gram language models were built for the words and full 4-gram models for the morphs (ignoring 3- and 4-grams found only once). No language model pruning was applied in our experiments.

5.2.3 Pronunciation and context dependency models

Simple grapheme-to-phoneme rules [19] and an exception list were used to generate word- and morph-to-phoneme mappings. The <w> symbols in the morph-based models were mapped to optional silences (similar to the 'short pause' (sp) model in [26]), while in the case of word-based models optional silences were added to the end of each word. The pronunciation models were further processed by applying triphone context expansion, as shown in equation (1) below. This includes not only the inter-word dependencies but the cross-word or cross-morph context dependencies as well, taking the optional inter-word silences into consideration. As was already mentioned in Section 4, phonological co-articulations were not considered explicitly by this pronunciation model.

5.3 The Recognition Network and the Decoder

The final step was the creation of a triphone level WFST (Weighted Finite State Transducer) [16] recognition network:

$$wred(fact(compact(C \circ S \circ compact(det(L \circ G)')))), \tag{1}$$

where the capital letters are transducers and the others are operators. This process commences with the composition and determinization of the language model (G) and the pronunciation model (L), then a suboptimal minimization process is applied. The optional silences are replaced with null transitions and silence models using the (S) transducer. Next, the context expansion is performed using the (C) transducer, then the network is minimized, factorized, and the weights are redistributed, resulting in a stochastic transducer suitable for a WFST decoder.

The above-described networks return word or morph sequences during the decoding process. Hence, a special operator has to be inserted in the computational process to obtain a phone sequence as output when the large vocabulary recognizer is used in phone recognition mode:

$$wred(fact(compact(C \circ proj(S \circ compact(det(L \circ G)'))))), \tag{2}$$

Table 3: Phone error rate (PER) results in [%]

	Training text					
Acoustic	NM (17.5K)		AM (1.4M)		SM (11.5M)	
Models	Word	Morph	Word	Morph	Word	Morph
triphone SI	26.9	23.9	13.7	14.2	13.7	14
triphone SA	19	13.2	6.7	6.5	5.8	6
triphone SD	14.6	7.4	3	3.1	2.5	2.7

where the projection operator copies the input labels of the silence model-substituted LG model in its output labels. All the operations were performed with "Mtool"* WFST building tool.

In the tests one-pass recognition was performed using the WFST decoder called VOXerver.* The tests were run on a Core 2 Quad processor at 2.67 GHz with 16 Gbytes of RAM. The RTF (Real Time Factor) of the morph- and the corresponding word-based system were adjusted so as to be nearly equal using standard pruning techniques. All tests ran in real-time (RTF<1) except for those that were performed with speaker independent acoustic models. In the following tests by the term relative improvement we mean the following:

$$Relative\,Improvement = \frac{ER_{reference} - ER_{new}}{ER_{reference}} * 100\% \qquad (3)$$

5.4 Results and Discussion

In this section, large vocabulary experimental results will be presented. First the phone error results (PERs) of section 4 are compared to PERs of large vocabulary recognizers, then word and letter error rates (WERs and LERs, respectively) are presented. After that, the morph-based improvements will be investigated. While WER and LER were calculated from the standard output of the large vocabulary decoder, special recognition networks had to be built to determine PER (see Section 5.3).

The primary aim of these experiments was to improve the phone recognition accuracy scores by utilizing higher level language models. As can be seen in Table 3, for most configurations we managed to outperform the phone-bigram models (the best error rates reported in Table 1 were around 14%). The improvement is especially good if there is a large training corpus available. By using large vocabulary language models even the distortion caused by poor acoustic modeling can be compensated for. Namely, the phone error rate that can be obtained with the combination of the largest language model and the general-purpose acoustic model is roughly the same as the scores of the best phone-bigram recognizers (~14%). The

*These tools were developed at AITIA International, Inc.

Table 4: Letter error rate (LER) results in [%]

	Training text					
Acoustic	NM (17.5K)		AM (1.4M)		SM (11.5M)	
Models	Word	Morph	Word	Morph	Word	Morph
triphone SI	29.1	25.5	14.3	14.4	14.3	14.4
triphone SA	21.7	15	7	6.3	6.1	5.8
triphone SD	17.7	9.9	3.3	2.8	2.5	2.4

Table 5: Word error rate (WER) results in [%]

	Training text					
Acoustic	NM (17.5K)		AM (1.4M)		SM (11.5M)	
Models	Word	Morph	Word	Morph	Word	Morph
triphone SI	68.3	61.9	37.1	37.4	36	36.8
triphone SA	65	50	24.9	22.6	21.6	20.8
triphone SD	61.6	41.6	17.5	14.2	13.4	12

only case where we got worse accuracies is when both the language and acoustic models were severely under-trained.

Large vocabulary speech recognizers are commonly characterized by their word and letter error rates. WER is the most widely applied way of evaluation, however, the LER provides a more realistic error measure in the case of morphologically rich languages. In our LER calculation the white spaces between words were modeled by a dedicated letter. Tables 4 and 5 summarize the word- and morph-based recognition results that were measured applying various acoustic models and training text corpora of various sizes.

The results clearly reveal the advantage of using a task-specific acoustic model. The best recognition results were attained with the model trained specially for the audiobook task. However, in many cases the available transcribed audio data is insufficient for training a new model. In such a case adapting the speaker independent model using the speaker-specific database – even in an unsupervised manner – provides a reasonable alternative.

A good match between the training and test sets is important, but it is not crucial for effective language modeling. As the results suggest, collecting large amounts of textual data is more rewarding than applying a well-matched but under-resourced database for the language model training. By combining the most elaborate models, the WER value was cut to 12%. To the best of our knowledge, this is the lowest WER reported on a Hungarian LVCSR task.

Having results with different language and acoustic models provides a great opportunity to investigate their impact on morph-based recognition improvements.

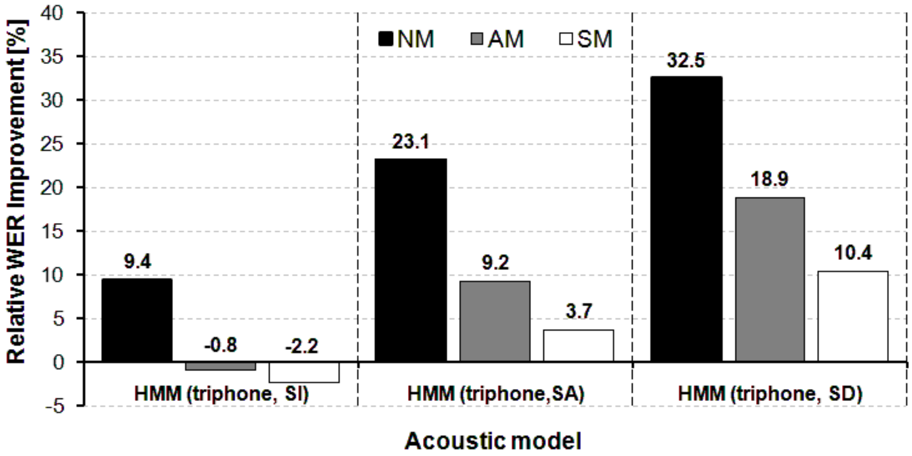

Figure 3: Relative improvements of WER due to change-over to morph-based language modeling. Improvement rates are measured using various acoustic models and training text corpora.

A commonly used metric for this is to measure the relative reduction of WER due to switching from word- to morph-based modeling (Fig 3). Looking at Figure 3 it can be seen that the fewer resources we have for training a language model the more beneficial is the morph-based approach. Furthermore, another useful conclusion is that by using a better matching acoustic model a higher improvement can be achieved. However, this also can turn into a deterioration in performance when the acoustic model is weakly-matched. These conclusions are in accord with the findings of [20].

6 Conclusions

In this paper we investigated the impact of various acoustical and language modeling techniques on the speech recognition accuracy measured on an audiobook recording. In such a near-ideal task the effect of disturbing acoustic factors (background noise, sloppy articulation, etc.) is almost negligible, thus the presented results can serve as baseline for the corresponding technologies in Hungarian continuous speech recognition. Despite the idealistic acoustic conditions, the task still represented various challenges in the language modeling step, due to the thematic and stylistic richness of the literary work the audiobook was based on.

First pure phone recognition experiments were presented, and the results clearly reflected that – at the acoustic level – the recognition of audiobooks is indeed much easier than that of a more realistic task. Our results with ANN-based acoustic models accord well with the similar international studies which indicate that these methods can attain similar or even better performance than the classic HMM-based triphone modeling technology. We also reinforced our earlier findings that triphone models do not require an explicit handling of phonological coarticulation during pronunciation modeling, while monophone models do.

In the large vocabulary recognition experiments we first measured phone recognition errors, but now using morph- and word-based language models instead of simple phone-bigrams. Since the phone-level output of our LVCSR system is not directly accessible, we had to build special recognition networks for this task. As expected, most of the large vocabulary supported recognition configurations outperformed the phone-based systems. We found that the larger the size of the training set, the higher the accuracy scores were. Comparing word- and morph-based phone error rates suggests that morphs are especially useful when the amount of the available training data is very limited. For instance, in the case of the novel-matched corpus the baseline phone-bigram phone error rate (∼14%) obtained in Section 4 could be improved only by introducing morph-based language modeling. On the contrary, when a large corpus is available, well-trained word-based language models may outperform morph-based ones in terms of PER, due to the ambiguous phonetic transcriptions at morph boundaries.

Though word-based recognizers may have higher phone-level accuracies, the morph-based configurations made consistently fewer letter errors. The quality of the recognized word sequence is closely related to the letter error rate, hence a morph-based recognizer is usually a better choice for large vocabulary tasks in Hungarian. Despite the advantages of LER, the word error rate is a more widely accepted metric. Therefore we used the WER to express the benefit of switching from word- to morph-based recognition. This improvement is especially good if the acoustic model suitably matches the recognition task, or when only a small corpus is available for language model training.

Comparing the error rates in this study to some of our earlier experimental results [20] we can get an impression of how acoustic factors of speech can degrade recognition performance. On a spontaneous speech task, which was recorded in a noisy environment and was both trained and evaluated with multiple speakers, roughly 50% WER was measured. In the case of press conference speeches – thanks to the higher signal-to-noise ratio and more planned speech production – the error rate dropped to 30%. While in near-ideal conditions WER can come close to 10%, as we see in this paper. Although these results cannot be directly compared due to differences in language and acoustic model sets, the tendency clearly shows that still much research have to be done to overcome the issues of real-life recognition tasks.

Numerous technologies have been investigated in this study, but there is still space for further development. In the future we would like to combine discriminative training methods and triphone acoustic models, since both approaches resulted in a consistent phone error rate improvement. A further important research direction would be to find a way of combining the advantages of the neural net-based and the triphone technologies in acoustic modeling. For speaker adaptation, it would be worth trying a supervised MAP adaptation of the speaker-independent acoustic model, since a good quality transcription of the audiobook is available. We expect that with this technique the performance of the adapted acoustic model could get closer to the speaker-specifically trained one.

References

[1] Aradilla, G., Bourlard, H., Magimai-Doss, M. Using KL-based Acoustic Models in a Large Vocabulary Recognition Task. Proceedings of Interspeech 2008: 928-931.

[2] Bourlard, B., Morgan, N. Connectionist Speech Recognition - A Hybrid Approach. Kluwer Academic, 1994.

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

[4] Creutz, M. and Lagus, K. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Comp. and Inf. Sci., report A81, HUT, March 2005.

[5] He, X., Deng, L. Discriminative Learning for Speech Recognition: Theory and Practice Morgan & Claypool, 2008.

[6] Hermansky, H., Ellis, D., Sharma, S. Tandem connectionist feature extraction for conventional HMM systems. Proceedings of ICASSP 2000: 1635-1638.

[7] Hirsimaki, T. and Kurimo, M. Decoder issues in unlimited Finnish speech recognition. Proceedings of the Nordic Signal Processing Symposium *NORSIG 2004*, Espoo, Finland, 2004.

[8] Huang, X., Acero, A., Hon, H.-W. Spoken Language Processing. Prentice Hall, 2001.

[9] Leggetter, C.J. and Woodland, P.C. Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. Proc. ARPA Spoken Language Technology Workshop, 1995.

[10] Lippmann, R. P. Speech Recognition by Machines and Humans. Speech Communication, 22(1): 1-15, 1997.

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

[12] Mihajlik P., Tatai P. Automatic phonetic transcription for Hungarian (In Hungarian). Beszédkutatás 2001: 172-185.

[13] Mihajlik P., Tüske Z., Tarján B., Németh B. and Fegyó T. Improved recognition of spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task. in IEEE Transactions on Audio, Speech, and Language Processing Vol. 18, Issue 6, pp. 1588-1600, 2010.

[14] Mihajlik P. Coarticulation Modeling in Automatic Speech Recognition for Hungarian (In Hungarian). Proceedings of MSZNY 2006, pp. 231-242.

[15] Mohamed, A.-R., Dahl, G., Hinton, G. Deep Belief Networks for Phone Recognition. Proceedings of NIPS 2009.

[16] Mohri, M., Pereira, F. and Riley, M. Weighted Finite-State Transducers in speech Recognition. Computer Speech and Language, 16(1), pp. 69-88, 2002.

[17] Siniscalchi, S. M., Schwartz, P., Lee, C.-H. High-Accuracy Phone Recognition By Combining High-Performance Lattice Generation and Knowledge-Based Rescoring. Proceedings of ICASSP 2007, pp. 869-872.

[18] Stolcke, A. J.T. SRILM – an extensible language modeling toolkit. Proc. Intl. Conf. on Spoken Language Processing, pp. 901–904, Denver, 2002.

[19] 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.

[20] Tarján B. and Mihajlik P. On Morph Based LVCSR Improvements. in Proc. of the 2nd Int. Workshop on Spoken Language Technologies for Under-resourced Languages, pp. 10–15, 2010.

[21] Tóth L., Frankel, J., Gosztolya G., King, S. Cross-lingual Portability of MLP-Based Tandem Features - A Case Study for English and Hungarian. Proceedings of Interspeech 2008: 2695-2698.

[22] Tóth Sz. L., Sztahó D., Vicsi K. Speech Emotion Perception by Human and Machine Proceedings of COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007.

[23] Vicsi K., Tóth L., Kocsor A., Gordos G., Csirik J. MTBA - Hungarian Telephone Speech Database (In Hungarian). Híradástechnika, Vol. LVII, No.8, pp. 35-43, 2002.

[24] Vicsi K., Kocsor A., Teleki Cs., Tóth L. Speech Database for Office Computer Environments (In Hungarian) Proceedings of MSZNY 2004 (2004), pp. 315-318.

[25] Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A. Effect of speaking style on LVCSR performance. Proceedings of ICSLP 1996: 16-19.

[26] Young, S., Ollason, D., Valtchev, V. and Woodland, P. The HTK book. (for HTK version 3.4), March 2009. http://htk.eng.cam.ac.uk

[27] Zhao, Y. Speech-Recognition Technology in Health Care and Special-Needs Assistance. IEEE Signal Processing Magazine, 26(3): 87-90, 2009.

[28] Hungarian Electronic Library (Magyar Elektronikus Könyvtár). http://www.mek.oszk.hu

[29] Digital Academy of Literature (Digitális Irodalmi Akadémia). http://www.irodalmiakademia.hu

[30] Electronic Archive of Periodicals (Elektronikus Periodika Archívum és Adatbázis). http://epa.oszk.hu