Int J Speech Technol (2006) 9: 121–131
DOI 10.1007/s10772-008-9008-2

Investigating the robustness of a Hungarian medical dictation system under various conditions

András Bánhalmi · Dénes Paczolay · László Tóth · András Kocsor

Received: 11 April 2007 / Accepted: 9 October 2008 / Published online: 2 December 2008

Abstract This paper examines the susceptibility of a dictation system to various types of mismatches between the training and testing conditions. With these experiments we intend to find the best training configuration for the system and also to evaluate the efficiency of the speaker adaptation algorithm we use. The paper first presents the components of the dictation system, and then describes a set of training and recognition experiments where we vary the microphones and create gender-dependent and speaker-dependent models. In each case we examine how much the recognition performance can be improved further by speaker adaptation. We conclude that the best and most reliable scores can be obtained by using gender-dependent phone models in combination with speaker adaptation. Speaker adaptation results in great improvements in almost every case. However, our results do not confirm the assumption that the use of one microphone is better than the use of several.

Keywords Dictation system · Medical dictation · Speaker adaptation · Hidden Markov Model

This work was supported by the National Grant IKTA-056/2003.

A. Bánhalmi · D. Paczolay · L. Tóth (✉) · A. Kocsor
Hungarian Academy of Sciences, Budapest, Hungary
e-mail: tothl@inf.u-szeged.hu

1 Introduction

There is common agreement in the speech recognition society that "there is no data like more data". That is, theoretically, the easiest and safest way of increasing the robustness of a speech recognition system is to collect training data from all possible test conditions. In practice, however, the variability of speech samples due to the differences among speakers, environments, recording conditions and so on is so large that it is impossible to cover every single combination. Hence, on one hand the limited amount of training data has to be designed and selected carefully. On the other hand, such methods as data normalization and adaptation may bring significant improvements in the system's performance. In this paper we first describe the building blocks of a Hungarian medical dictation system, and then present a set of experiments we conducted in order to find the best training configuration of the system and in particular to see how much improvement can be obtained by speaker adaptation.

The experiments presented here were all performed within the framework of a medical dictation project. This project was initiated by two university departments with financial support from a national fund, and its main goal was to create the first Hungarian continuous dictation system. In the first part of the paper we explain the motivations behind the project, and then we describe all the modules of our system—the acoustic model, the language model, the user interface and the adaptation algorithm—in detail. In the

experimental part we try to find the optimal training configuration for the system. For this we examine whether a significant increase in performance can be obtained by insisting on employing one specific microphone during both training and testing, instead of allowing the usage of various microphones with supposedly quite different transfer characteristics. We also evaluate the system's behavior after training separate, gender-specific phone models for the male and female speakers.

In the past few years our team has participated in several Hungarian speech recognition projects, but in these experiments we have so far focused only on building speaker independent models. In the framework of the medical dictation project we came to realize that in the medical report dictation task our speaker-independent models do not perform well enough. Since this task allows the application of speaker adaptation during recognition, our other main goal with the experiments described here was to examine how much the performance of our system could be improved by applying adaptation techniques to our speaker independent models. For this purpose we applied speaker adaptation after the baseline tests in each training situation to see whether it could bring further improvements. Finally, we also performed a speaker-dependent training experiment where we trained the model on the voice of one speaker and then adapted it to the voice of another. We think that such an experiment can show most clearly how useful speaker adaptation can be.

2 The Hungarian medical dictation project

At the present time there exists no general-purpose large vocabulary continuous speech recognizer (LVCSR) for the Hungarian language. Among the university publications even papers that deal with continuous speech recognition are hard to find, and these present results for restricted vocabularies only (Szarvas and Furui 2002). Although on the industrial side Philips have adapted its SpeechMagic system to two special domains in Hungarian, it is sold at a price that is affordable for only the largest institutes (Medisoft 2004). The experts usually mention two reasons for the lack of Hungarian LVCSR systems. First, there are no sufficiently large, publicly available speech databases that would allow the training of reliable phone models. The second reason is the difficulties of language modeling due to the highly agglutinative nature of Hungarian.

In 2004 the Research Group on Artificial Intelligence of the University of Szeged and the Laboratory of Speech Acoustics of the Budapest University of Technology and Economics started a project with the aim of collecting and/or creating the basic resources needed for the construction of a continuous dictation system. The project lasted for three years (2004–2006), and was financially supported by the national fund IKTA-056/2003. As regards acoustic modeling, the project included the collection and annotation of a large speech corpus of phonetically rich sentences. For language modeling, we restricted the target domain to the dictation of certain types of medical reports. Although this clearly led to a significant reduction compared to the original, general dictation task, we chose this application area with the intent of assessing the capabilities of our acoustic and language modeling technologies. Depending on the findings, later we hope to extend the system to more general dictation domains. This is why the language resources were chosen to be domain-specific, while the acoustic database contains quite general, domain-independent recordings.

Although both participating teams used the same speech database to train their acoustic models, they focused on two different dictation tasks and experimented with their own acoustic and language modeling technologies. Our team in Szeged focused on the task of the dictation of thyroid scintigraphy medical reports, while the Budapest team dealt with gastroenterology reports. This paper describes the recognition system and development efforts of the Szeged team only.

3 Components of the medical dictation system

3.1 Acoustic modeling

Hungarian is a Finno-Ugric language, so it is one of the few modern European languages that do not belong to the Indo-European language family. Owing to this, there are several significant differences between the phonetics of Hungarian and English (Szende 1999). Their consonant sets are relatively similar, the biggest mismatches being the dental fricatives of English and the palatal affricates of Hungarian, which

are missing from the other language. However, there are also several allophonic differences in those consonants that are present in both languages (for example, voiceless stop consonants are never aspirated in Hungarian). There are much bigger differences in the vowel systems. Even the similar monophthongs take slightly different positions in the vowel triangle, while some of them are missing from the other language (e.g. /æ/ from Hungarian, /ø/ and /y/ from English). But even more importantly, in Hungarian there are no diphthongs (apart from dialects and sloppy speech) and unstressed vowels do not get reduced (or only to a much lesser extent than in English). Probably the most exotic feature of the phonology of Hungarian is that most consonants and vowels have a long and a short variant, and their duration acts as a distinctive feature.

Apart from this little peculiarity, the conventional acoustic modelling techniques such as the Hidden Markov Model (HMM) (Huang et al. 2001) are readily applicable to the recognition of Hungarian. In the experiments reported here we used a quite standard HMM decoder implementation. This system works with the also conventional mel-frequency cepstral coefficient (MFCC) features (Huang et al. 2001). More precisely, 13 coefficients are extracted from each 25 msec frame, along with their Δ and $\Delta\Delta$ values, at a frame rate of 100 frames per sec. The phone models applied have the usual 3-state left-to-right topology. As the duration feature in the vocabulary of our specific dictation task seemed to have no discriminative role, most of the long/short consonant labels were fused, and this way we worked with just 44 phone classes. One phone model was associated with each of these classes, that is we applied monophone modeling and no context-dependent models were tested in the system. The decoder built on these HMM phone models performs a combination of Viterbi and multi-stack decoding. To speed up the process it contains several built-in pruning criteria. First, it applies beam pruning, so just the hypotheses with a score no worse than the best score minus a threshold are kept. Second, the number of hypotheses extended at each time point is restricted, which corresponds to multi-stack decoding with a stack size constraint. The maximal evaluated phone duration can also be restricted. Normally on a typical PC the decoder runs faster than real-time on our dictation task.

3.2 Language modeling

A special difficulty of creating language models for Hungarian is the highly agglutinative nature of the language. Thus in a large vocabulary modeling task the application of a morphologic analyzer/generator algorithm seems inevitable. First, simply listing and storing all the possible word forms would be nearly impossible (an average noun can have about 700 inflected forms). Second, if we simply handled all these inflected forms as different words, then achieving a certain coverage rate in Hungarian would require a text about 5 times bigger than that in German and 20 times bigger than that in English (Németh and Zainkó 2001). Consequently, the training of conventional n-gram language models would require significantly larger corpora in Hungarian than in English, or even in German. A possible solution might be to train the n-grams over morphemes instead of word forms, but then again the handling of morphology would be necessary.

Though quite good morphological tools now exist for Hungarian, in the first experiments with our dictation system we preferred to avoid the complications with morphology. The restricted vocabulary is one of the reasons why we chose the medical dictation task. As we mentioned earlier, our analysis revealed that the thyroid gland medical reports we examined contained only about 2500 different word forms. Although these many words could be easily managed even by a simple list ('linear lexicon'), we organized them into a lexical tree (Fig. 1) where the common prefixes of the lexical

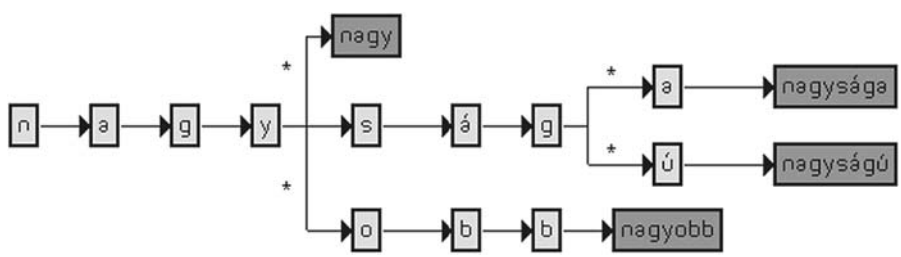

Fig. 1 Prefix tree for some Hungarian words. At the points labelled by *asterisks* the grammar model can generate the exact n-gram probability of the given word

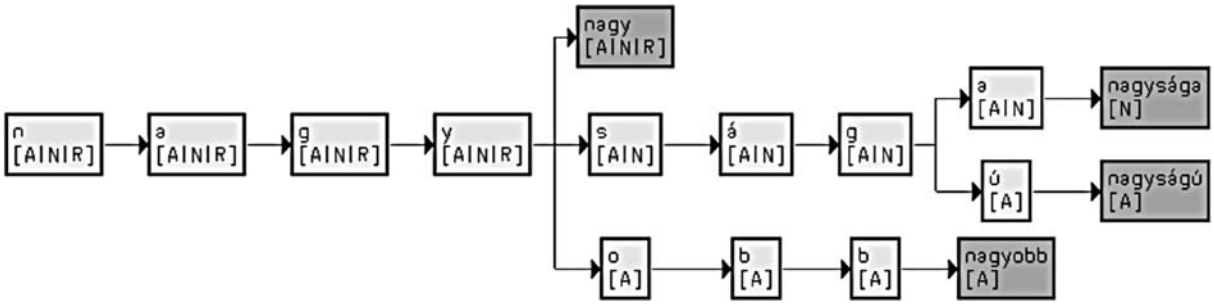

Fig. 2 Prefix tree for some Hungarian words with their POS code. At the branches of the tree the grammar model can generate the probability of the word according to the word n-gram and according to the class n-gram

entries are shared. Apart from storage reduction advantages, this representation also speeds up decoding as it eliminates redundant acoustic evaluations (Huang et al. 2001). The prefix tree representation is probably even more useful for agglutinative languages than for English because of the many inflected forms of the same stem.

The limited size of the vocabulary and the highly restricted (i.e. low-perplexity) nature of the sentences used in the reports allowed us to create very efficient n-grams. Moreover, we did not really have to worry about out-of-vocabulary words, since we had all the reports from the previous six years, so the risk of facing unknown words during usage seemed minimal. The system currently applies 3-grams by default, but it is able to 'back off' to smaller n-grams (in the worse case to a small ϵ constant) when necessary. During the evaluation of the n-grams the system applies a language model lookahead technique (Huang et al. 2001). This means that the language model returns its scores as soon as possible, not just at word endings. For this purpose the lexical trees get factored, so that when several words share a common prefix, a combination of their probabilities is associated with that prefix. The probability of a word prefix given by the grammar model will be this combined value, which will change at every node of the prefix tree. These techniques allow a more efficient pruning of the search space (Bánhalmi et al. 2005).

Besides word n-grams we also experimented with constructing class n-grams. To do this the words were grouped into classes according to their parts-of-speech category. The words were categorized using the POS tagger software developed at our university (Kuba et al. 2004). This software associates one or more MSD (morpho-syntactic description) code (Erjavec and Monachini 1997) with the words, and we constructed the class n-grams over these codes (Fig. 2). Although we utilized only the first letter of the MSD code—which is practically the POS code—the MSD code would allow the construction of more sophisticated word classes as well. With the help of the class n-grams, the language model can be made more robust in those cases when the word n-gram encounters an unknown word, so it practically performs a kind of language model smoothing. In previous experiments we found that the application of the language model lookahead technique and class n-grams brought about a 30% decrease in the word error rate when it was applied in combination with our HMM-based fast decoder (Bánhalmi et al. 2005).

A high performance improvement can be achieved using assimilation rules when concatenating word models (Kocsor et al. 2006). This feature has also been incorporated into our grammar model. Because of the many possible utterances of a word, the grammar model is able to search among the shared common suffixes as well, so the model is not necessary a prefix tree, but it may become a directed graph.

3.3 User interface

The GUI was really designed with a view to serving many users on the same computer. But also the GUI was intended to combine simplicity with good functionality. Just a microphone and a text editor (Microsoft Word, or any word processor package) are needed for dictating medical reports.

Every user has one or more profiles containing all the special information characterizing his or her

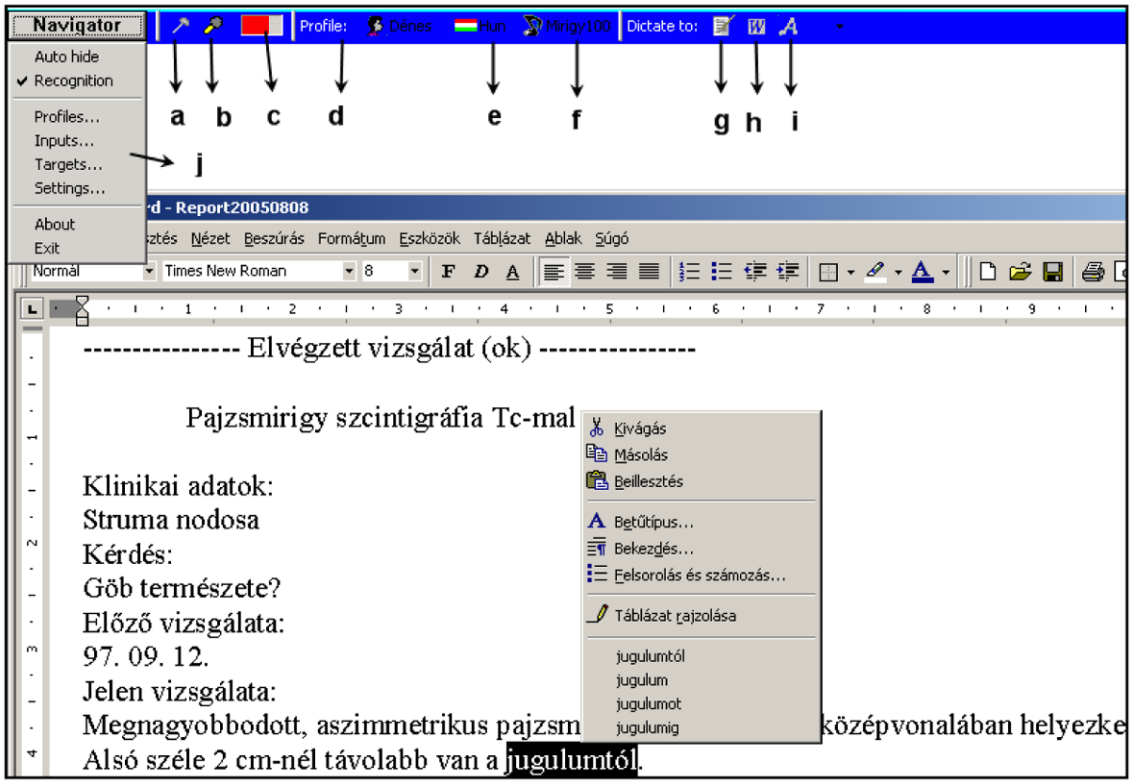

Fig. 3 Functions of the graphical user interface: (**a**) Enable or disable auto hiding of the main toolbar. (**b**) Start or stop the recognition procedure. The user can suspend the dictation at any time, and can continue later. (**c**) Volume display bar. The volume of the microphone input can be checked here. (**d**) Choosing a specific user. The user can be selected from the list of existing users. (**e**) Choosing the actual language. A language assigned to the current user can be chosen from a listbox. (**f**) Choosing the actual grammar. An available grammar can be chosen with just one click. (**g**) Selecting the internal text editor. The output of the dictation will be typed into this internal smart text editor. (**h**) Selecting the Microsoft Word plugin for output. (**i**) Selecting the window of the active application. With this function the user can dictate into any MS-Windows based application like MS Excel or MS Outlook. (**j**) the main menu for managing the user profiles. All the above-mentioned functions are accessible here

voice for a given language and vocabulary. The language models and the acoustic core modules can be installed separately. These models will be adapted to the characteristics of the users. The user interface also has a toolbar at the top of the desktop. Using the toolbar all the main functionalities related to the initial parameter settings can be accessed, like choosing a specific user, choosing the actual task and selecting the output window (Fig. 3). Other functionalities can only be accessed from the actual output text editor. The most important of these features is that the user can ask the speech recognition system for other possible variants of the recognized sentences if the recognized word or sentence is somehow incorrect.

4 The adaptation algorithm

The need for speaker adaptation methods in speech recognition applications arises in those tasks where we wish to achieve a speaker-dependent recognition performance, but there is no opportunity to perform a separate user-specific training phase. In such cases the development of a speaker-dependent model for each speaker is practically impossible, because the required large amount of speaker-specific training data is unavailable or difficult to acquire.

The two main approaches for improving the performance of a speaker-independent model are the transformation of the incoming feature vectors (by methods like VTLN or CMN) and the adaptation of the parameters of the statistical acoustic models. In classical

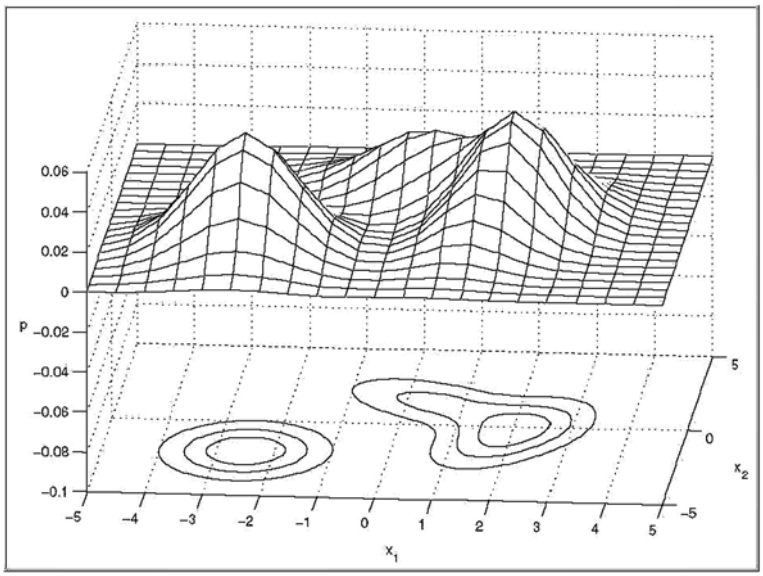

Fig. 4 Gaussian mixture model with three mixtures in two dimensions

HMM-based systems various speaker adaptation techniques have been used with success. These techniques are based on the fine-tuning of the parameters of the speaker-independent system to the given speaker in order to maximize the likelihood of the adaptation data of the new speaker.

In general, adaptation can be applied with three strategies: batch adaptation, self-adaptation and on-line adaptation. In the case of batch adaptation the adaptation is performed after all the adaptation data has been collected, so this is an off-line method. Self-adaptation is performed on the testing data at run-time. As this method is unsupervised, the recognition errors and the faulty transcripts pose a special problem. Various conditions were proposed earlier for filtering the words to be retained for the adaptation step (Matsui and Furui 1996; Homma and Sagayama 1997; Wessel 2002). The on-line (or incremental) adaptation technique changes the parameters of the statistical model only after a block of adaptation data has been processed, and this block of data is then thrown away, so this method is a trade-off between the first two techniques.

Computationally, two main approaches have been proposed in the literature for the adaptation of HMM parameters. The first is the maximum likelihood (ML)-based framework containing the maximum likelihood linear regression (MLLR) approach (Leggetter and Woodland 1995), the maximum likelihood stochastic matching (SM) approach (Sankar and Lee 1996) and the constrained transformation approach (Digalakis et al. 1995; Diakoloukas and Digalakis 1999). The other main group of techniques are based on the maximum a posteriori (MAP) formulation (Gauvain and Lee 1994), where the forward and backward probabilities are not fully computed and not all the HMM parameters are re-estimated, but the path with the maximal probability is determined, and just the parameters belonging to this path are re-estimated. As the speech decoding step in most LVCSR systems work in the same way (i.e. using Viteri search), this technique may work more reliably than MLLR-like techniques.

All the adaptation algorithms that are usually applied to HMMs are essentially based on the same idea, that is they adjust the parameters of the speaker independent HMMs so that the new values are more suitable for describing the speech of the actual speaker. Thus the tuned model is closer to a specific, speaker dependent model.

The HMM phone models consists of states with emission probabilities described by Gaussian mixtures, and transition probabilities between the states (Fig. 4). It was found that varying the transition parameters and the Gaussian mixture weight parameters have little effect on the performance of the recognition, and the most important parameters are the mean values of the Gaussian mixtures (Bourlard et al. 1996; Rozzi 1991). Hence, the adaptation process usually addresses only these latter parameters.

In order to re-estimate the Gaussian components of the HMMs our system applies the maximum a posteriori (MAP) adaptation method. The data used during the adaptation step is extracted on-line, embedded into the recognition process. This way the adaptation step is not separated algorithmically from the recognition step, but instead it operates just like the continuous recognizer. More precisely, the adaptation is performed as follows. The Viterbi algorithm performed during recognition produces a series of the most probable HMM phone models, the most probable state sequence within these models, and also the sequence of the Gaussian components belonging to the states. Having obtained the state and Gauss component index associated with each input feature vector, the mean of the given Gaussian component is updated using the MAP formulation (Thelen 1996):

$$\mu_{new} = \frac{N}{N+\alpha} m_{obs,ML} + \frac{\alpha}{N+\alpha} \mu_0, \tag{1}$$

where

$$m_{obs,ML} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2}$$

Here the parameter N represents the number of examples for the given mixture component (x_i), while the parameter α controls the speed of varying the mean of the mixture. These formulae practically perform a linear regression from the speaker-independent model to the speaker-dependent model.

In practice we apply a recursive version of the above formulae, which allows a continuous, incremental adaptation:

$$\mu_{d,N+1} = \frac{x_{N+1} + (N+\alpha) \cdot \mu_{d,N}}{N+1+\alpha}. \tag{3}$$

Our adaptation method can be used both for supervised adaptation and for unsupervised adaptation. When using it for unsupervised adaptation, the transcript of the spoken sentence is unknown. To get the transcript, the speech recognizer stores a derivation tree for the hypotheses. When a common ancestor becomes available, then its transcript and the corresponding speech signal can be used as the input for supervised adaptation.

5 Experiments

Our aim with the experiments was to find out how much the recognition performance depends on the gender of the train/test speakers, on the person who tests the system, and on the microphone being used. Obviously, we also wanted to see how much of the error caused by gender, speaker or microphone dependency could be reduced by speaker adaptation. To understand these relationships better, we varied the contents of the training database. For speaker and gender-independent training we used the MRBA corpus (Vicsi et al. 2004), which contains recordings from 100 speakers, both men and women. Then, in order to test the gender-dependency of the system, we created separate models for the male and female voices using the same corpus. To be able to perform speaker-dependent training as well, we recorded a longer speech item from one specific male and female speaker. Finally, a very special property of the MRBA corpus is that its full content was recorded with two microphones in parallel. One of these, which we will refer to as the 'primary' microphone was always the same during the recordings, while for the 'secondary' recordings a different microphone was used for each speaker. This fact allowed us to examine how much the recognizer's performance depends on the microphone used.

5.1 Parameters of the Corpora

The MRBA Corpus contains speech samples from 100 speakers, namely 26 women and 74 men. The age of the speakers is between 13 and 72 but, as Fig. 5 shows, the speakers are mostly from the most active computer-using generations. The speech signals were recorded and stored at a sampling rate of 16000 Hz in 16-bit quality (the same sampling rate and quality the dictation system operates at). Each speaker uttered 12 long sentences (16 words per sentence on average) and 12 phonetically rich words. The database contains a total of about 10,800 words (85,300 phonemes) in about 100 minutes of recorded speech material. As mentioned above, each utterance was recorded via two microphones (primary and secondary) simultaneously. The primary microphone was fixed, while the secondary was varied for each speaker (a variety of cheap computer microphones were tested). in the tables below the sentences recorded by the primary microphone

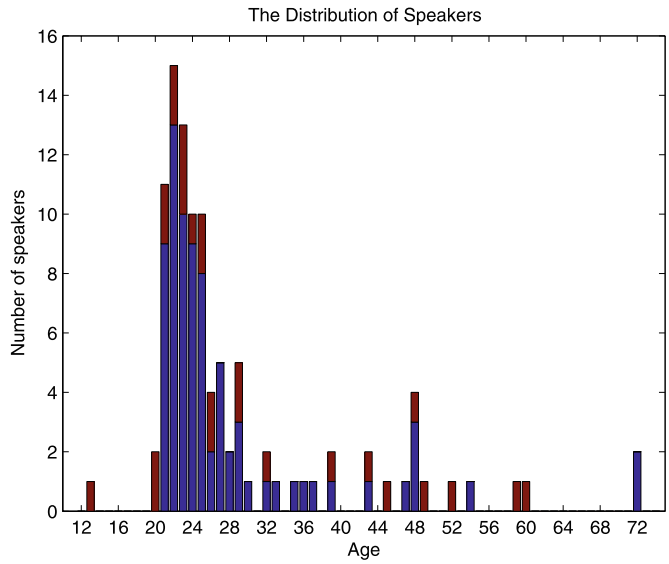

Fig. 5 The age distribution of the speakers in the MRBA corpus

Table 1 Properties of the databases

	Train	Adapt	Test	
#speakers	100	2×1	5×1	5×1
#sentences	1,200	2×240	5×86	1,380
#words	1,200	2×240	–	–
total #words	10,800	$2 \times 2{,}200$	5×613	$5 \times 7{,}000$
total #phonemes	85,300	$15{,}500 + 15{,}600$	n/a	n/a
total length (min)	100	$21 + 23$	5×6	5×15

are denoted by an '-r' tag, while those recorded by the secondary microphone set are denoted by an '-s' tag.

For the **speaker-specific training** experiments two persons (a male and a female) uttered 240/240 sentences and 240/240 words, which in total gave 2200/2200 words (15500/15600 phonemes) in 21/23 min. The speech signals were recorded and stored using the same technique mentioned above but with just the primary microphone.

As **adaptation and testing data** we recorded samples from 5 speakers—2 men and 3 women (in the following they will be referred to as M1, M2, W1, W2 and W3). The speech signals were recorded and stored using the same technique mentioned above. Each speaker spoke the same 86 sentences in 17 paragraphs (613 words) for adaptation and 20-20 different medical reports for test. The length of the speech signal recorded for adaptation purposes was about 6 minutes, and that for testing was about 15 minutes per speaker. Speakers denoted by M1 and W1 were the same as those who produced the speech material for speaker specific training, so this way we had the opportunity to measure the discrepancy between inter-speaker and cross-speaker training and testing.

Table 1 below gives a summary of the parameters of the train (general and speaker-specific), adaptation and test databases.

6 Results and discussion

In the first experiment we utilized all the training data of the secondary microphone part of the MRBA corpus to create just one set of phone models. The testing results for the five test persons (two males, three females) are shown in Table 2 below. The scores are all reasonable, except for one of the females, whose score is significantly worse than the rest. Also, the results for the males are somewhat better, which might be due to the 3:1 ratio of males to females in the training database.

As the next step we repeated the above experiment, but this time using the data recorded with the primary microphone for both training and testing. One would naively expect that using one specific microphone for recording all the data should improve the scores as it would remove the variance caused by the differences in the microphone transfer characteristics. Surprisingly, we found that all the results were worse (see Table 3). Currently we cannot satisfactorily explain this behavior, but our hypothesis is that by applying many different microphones in the 's' training set the phonetic models had a better generalization ability and hence became more robust than when using just the specific 'r' microphone during both training and testing.

We also attempted to improve both the microphone-specific and microphone-independent models by speaker adaptation. Table 4 summarizes the results that we obtained. Comparing the corresponding lines with those of Tables 2 and 3, we see that it brought a slight degradation for the male speakers, and a small and a huge improvement for the female speakers W3 and W1, respectively, and for W2 a huge improvement in one case, but a similarly huge decrease in the other. So altogether it seems that the adaptation process can have a beneficial effect, but it should be used with caution.

In the next experiment we separated the male and gender test data and trained a separate, gender-specific phone set on them. This is quite a common practice in phonetic modelling (Huang et al. 2001). The recognition results (see Table 5) then became much more balanced, and the scores of the problematic W1 speaker became more like the others. However, as a price, the results of the consistently good speakers (M1, M2, W3) fell. We think that this can be attributed to the fact that because of the separate training of the genders the training data for the models was practically halved. With a much larger training corpus the improvement due to the gender-specific training would probably be

Table 2 Results when using the mixed microphones (set 's') during training

Test database				
Men		Women		
M1-s	M2-s	W1-s	W2-s	W3-s
95,52%	97.64%	79.26%	91.30%	94.23%

Table 3 Results when using the specific 'r' microphone during training

Test database				
Men		Women		
M1-r	M2-r	W1-r	W2-r	W3-r
93.99%	97.00%	75.49%	72.49%	91.17%

Table 4 Results after speaker adaptation for both the 's' and 'r' sets

Train mic	Test database				
	Men		Women		
	M1-s	M2-s	W1-s	W2-s	W3-s
s	94.72%	98.36%	89.13%	87.39%	94.80%
	M1-r	M2-r	W1-r	W2-r	W3-r
r	92.71%	96.42%	95.61%	87.84%	93.02%

Table 5 Results with gender-dependent acoustic models

		Test database				
		Men		Women		
		M1	M2	W1	W2	W3
no adaptation	s. mic	94.35%	95.42%	84.88%	83.52%	85.13%
	r. mic	92.16%	92.21%	82.76%	80.89%	91.94%
after adaptation	s. mic	92.30%	93.00%	89.41%	85.09%	91.73%
	r. mic	90.29%	94.78%	94.51%	89.07%	93.58%

Table 6 Results of speaker-dependent training

	Test database				
	Men		Women		
	M1	M2	W1	W2	W3
no adaptation	89.34%	77.50%	83.72%	53.89%	51.04%
after adaptation	96.71%	93.70%	87.14%	81.77%	87.83%

much bigger than the drawback caused by the relative decrease in the amount of training data per model.

Finally we examined what kind of results could be obtained by training speaker-dependent models. For this we trained a model on the training data of speakers M1 and W1. Besides their own corresponding test data, we also tested the model of M1 on the data of M2, and the model of W1 on the test data of speakers W2 and W3. Then we repeated the testing after speaker adaptation. The results of this are shown in Table 6. We can clearly see that without adaptation the test results are not good already for the test data of the training speakers, and even worse for the other people. After adaptation all the scores improved quite significantly—in particular for those persons whose material was not used during training. Although, on average, the results after adaptation are still a bit worse than those scores obtained with speaker-independent training, we think that this is because of the much smaller size of the speaker-dependent training data set. The fact that there was a large increase even for the training speakers M1 and W1 indicates that adding more training material would be just as important as speaker adaptation itself. Nevertheless the marked increase in the scores for the other speakers clearly demonstrates the efficiency of the adaptation algorithm in the case where there is a speaker discrepancy between the train and test recordings.

7 Conclusions

This paper examined the susceptibility of a dictation system to various kinds of mismatches between the training and testing conditions. In our experiments we sought to find the best training configuration for the system and also to evaluate the efficiency of the speaker adaptation algorithm we use. Our studies did not justify any advantage of using one specific microphone during both training and testing. It seems that training using several microphones made the system more robust for the specific microphone as well. The gender-specific training decreased the variance of the results over the test persons, but the average of the scores became slightly worse. We think that this is because the training database is very small, and hence halving it by separating the data items according to gender can bring about such a degradation. With a much larger training set this fall in the average performance would presumably be much smaller, and the advantages of the gender-dependent models would become more apparent. The point that our training data is just barely enough is also indicated by the speaker dependent modelling experiment where the speaker adaptation algorithm brought a significant improvement even for the speaker whose voice was used to train the model. This improvement was apparently not due to the adaptation, but rather to the increase in the amount of training data. For the test persons different from the train persons the adaptation brought a good performance improvement, and the adaptation algorithm improved the scores in almost all the other training configurations as well. So our two main conclusions from the experiments is that the best setup to apply is gender-specific phone modelling in combination with speaker adaptation, and that we should also increase our training data set if we want to see more consistent and reliable behavior.

References

Bánhalmi, A., Kocsor, A., & Paczolay, D. (2005). Supporting a Hungarian dictation system with novel language models. In *Proc. of the 3rd Hungarian conf. on computational linguistics*, pp. 337–347, 2005 (in Hungarian).

Bourlard, H., Hermansky, H., & Morgan, N. (1996). Towards increasing speech recognition error rates. *Speech Communication*, *18*(3), 205–231.

Diakoloukas, V., & Digalakis, V. (1999). Maximum-likelihood stochastic-transformation adaptation of hidden Markov models. *IEEE Transactions on Speech Audio Processing*, *2*(7), 177–187.

Digalakis, V., Rtischev, D., & Neumeyer, L. (1995). Speaker adaptation using constrained reestimation of Gaussian mixtures. *IEEE Transactions on Speech Audio Processing*, *3*, 357–366.

Erjavec, T., & Monachini, M. (Ed.) (1997). Specification and notation for lexicon encoding. In *Copernicus project 106 "MULTEX-EAST"*, Work Package 1—Task 1.1, Deliverable D1.1F.

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291–298.

Homma, K. A., & Sagayama, S. (1997). Improved estimation of supervision in unsupervised speaker adaptation. In *Proc. of ICASSP-97*, pp. 1023–1026, 1997.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing*. New York: Prentice Hall.

Kocsor, A., Bánhalmi, A., & Paczolay, D. (2006). Methods of informatics and mathematics in a system for dictating thyroid gland medical reports. *Acta Agraria Kaposvariensis*, *10*(1), 113–128 (in Hungarian).

Kuba, A., Hócza, A., & Csirik, J. (2004). POS tagging of Hungarian with combined statistical and rule-based methods. In *Proc. of TSD 2004*, pp. 113–121, 2004.

Leggetter, C., & Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, *9*, 171–185.

Matsui, T., & Furui, S. (1996). N-best-based instantaneous speaker adaptation method for speech recognition. In *Proc. of ICSLP-96*, pp. 973–976, 1996.

Medisoft (2004). www.medisoftspeech.hu.

Németh, G., & Zainkó, C. (2001). Word unit based multilingual comparative analysis of text corpora. In *Proc. of eurospeech 2001*, pp. 2035–2038.

Rozzi, M. (1991). *Speaker adaptation in continuous speech recognition via estimation of correlated mean vectors*. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Sankar, A., & Lee, C. H. (1996). A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *3*(4), 190–202.

Szarvas, M., & Furui, S. (2002). Finite-state transducer based Hungarian LVCSR with explicit modeling of phonological changes. In *Proc. of ICSLP 2002*, pp. 1297–1300.

Szende, T. (1999). *Illustrations of the IPA: Hungarian, Handbook of the International Phonetic Association* (pp. 104–107). Cambridge: University Press.

Thelen, E. (1996). Long term on-line speaker adaptation for large vocabulary dictation. In *Proc. of IEEE ICSLP*, pp. 2139–2142, 1996.

Vicsi, K., Kocsor, A., Teleki, C., & Tóth, L. (2004). Hungarian speech database for computer-using environments in offices. In *Proc. of the 2nd Hungarian conf. on computational linguistics*, pp. 315–318, 2004 (in Hungarian).

Wessel, F. (2002). *Word posterior probabilities for large vocabulary continuous speech recognition*. PhD thesis, RWTH Aachen University.