Development of a Hungarian Medical Dictation System

András Bánhalmi, Dénes Paczolay, Lászó Tóth and András Kocsor†
Research Group on Artificial Intelligence
Hungarian Academy of Sciences and the University of Szeged
Aradi vértanúk tere 1, H-6720 Szeged
{banhalmi, pdenes, tothl, kocsor}@inf.u-szeged.hu

† Applied Intelligence Laboratory Ltd. and
Research Group on Artificial Intelligence NPC,
Petőfi Sgt. 43, H-6723 Szeged, Hungary

Keywords: speech recognition, dictation systems, 2D-cepstrum

Received: May 12, 2004

This paper reviews the current state of a Hungarian project which seeks to create a speech recognition system for the dictation of thyroid gland medical reports. First, we present the MRBA speech corpus that was assembled to support the training of general-purpose Hungarian speech recognition systems. Then we describe the processing of medical reports that were collected to help the creation of domain-specific language models. At the acoustic modelling level we experimented with two techniques – a conventional HMM one and an ANN-based solution – which are both briefly described in the paper. Finally, we present the language modelling methodology currently applied in the system, and round off with recognition results on test data taken from four speakers. The scores show that on a somewhat restricted sub-domain of the task we are able to produce word accuracies well over 95%.

Povzetek: Prispevek predstavlja pregled trenutnega stanja madžarskega projekta, ki skuša vzpostaviti sistem razpoznavanja govora za narekovanje zdravniških izvidov na temo žleze ščitnice.

1 Introduction: state of the art and goals of the project

Automating the dictation of texts is one of the main applications of speech recognition. Mainly because of the huge training corpora, the increased processor speeds and the refined search techniques dictation systems have reached such a level of sophistication that the commercial products now offer sufficiently good accuracy even for arbitrary normal-pace fluent speech [12]. Experience tells us, however, that for a really good performance it is still worth applying some tricks like an initial speaker enrollment process where the machine can adapt to the voice of the speaker, or the restriction of the dictation topic to some specific (e.g. medical or legal) domain. Such dictation systems already exist for the biggest languages, but the situation for those languages that can offer only a small market is not as good. For Hungarian at the present time there exists no general-purpose large vocabulary continuous speech recognizer (LVCSR). Among the university publications even papers that deal with continuous speech recognition are hard to find, and these give results only for restricted vocabularies [15]. Although on the industrial side Philips have adapted its SpeechMagic system to two special application domains in Hungarian, it is sold at a price that is affordable for only the largest institutes [9]. The experts usually cite two main 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 special difficulties of language modelling that arise due to the highly agglutinative nature of Hungarian.

In 2004 the Research Group on Artificial Intelligence at the University of Szeged and the Laboratory of Speech Acoustics of the Budapest University of Technology and Economics began a project with the aim of collecting and/or creating the basic resources needed for the construction of a continuous dictation system for Hungarian. The project lasted for three years (2004-2006), and was financially supported by the national fund IKTA-056/2003. For the acoustic modelling part, the project included the collection and annotation of a large speech corpus of phonetically rich sentences. As regards the language modelling part, we restricted the target domain to the dictation of some limited types of medical reports. Although this clearly led to a significant reduction compared to a general dictation task, we chose this application area with the intent of assessing the capabilities of our acoustic and language modelling 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 modelling technologies. The team at the University of Szeged focused on the task of the dictation of thyroid scintigraphy medical reports, while the Budapest team dealt with gastroenterology reports. This paper just describes the research and development efforts of the Szeged team. The interested reader can find a survey of the research done by the Laboratory of Speech Acoustics in [16].

2 Speech and language resources

In the first phase of the project we designed, assembled and annotated a speech database called the MRBA corpus (the abbreviation stands for the "Hungarian Reference Speech Database") [16]. Our goal was to create a database that allows the training of general-purpose dictation systems which run on personal computers in office environments and operate with continuous, read speech. The contents of the database were designed by the Laboratory of Speech Acoustics. As a starting point, they took a large (1.6 MB) text corpus and after automatic phonetic transcription they created phone, diphone and triphone statistics from it. Then they selected 1992 different sentences and 1992 different words in such a way that 98.8% of the most frequent diphones had at least one occurrence in them. These sentences and words were recorded from 332 speakers, each reading 12 sentences and 12 words. Thus all sentences and words have two recordings in the speech corpus. Both teams participated in the collection of the recordings, which was carried out in four big cities, mostly at universities labs, offices and home environments. In the database the ratio of male and female speakers is 57.5% to 42.5%. About one-third of the speakers were between 16 to 30 years of age, the rest being evenly distributed among the remaining age groups. Both home PCs and laptops were used to make the recordings, and the microphones and sound cards of course varied as well. The sound files were cleaned and annotated at the Laboratory of Speech Acoustics, while the Research Group on Artificial Intelligence manually segmented and labelled one third of the files at the phone level. This part of the corpus is intended to support the initialization of phone models prior to training on the whole corpus.

Besides the general-purpose MRBA corpus, we also collected recordings that are specific for the target domain, namely thyroid scintigraphy medical reports. From these recordings 20-20 reports read aloud by 4 persons were used as test data in the experiments reported here.

For the construction of the domain-specific language models, we got 9231 written medical reports from the Department of Nuclear Medicine of the University of Szeged. These thyroid scintigraphy reports were written and stored using various software packages that were employed at the department during 1998 to 2004. So first of all we had to convert all the reports to a common format, followed by several steps of routine error correction. Each report consists of 7 fields: header (name, ID number etc. of the patient), clinical observations, request of the referral doctor, a summary of previous examinations (if any), the findings of this examination, a one-sentence summary, and a signature. From the corpus we omitted the first and the last, person-specific fields, for the sake of personal data privacy. Then we discarded those reports that were incomplete such as those that had missing fields. This way only 8546 reports were kept, which, on average, contained 11 sentences and 6 words per sentence. The next step was to remove any typographical errors from the database, of which there were surprisingly many (the most frequent words occurred in 10-15 mistyped forms). A special problem was that of unifying the Latin terms, many of which are allowed to be written both with a Latin or a Hungarian spelling in medical texts (for example *therapia* vs. *terápia*). The abbreviations also had to be resolved. The corpus we got after these steps contained approximately 2500 different word forms (excluding numbers and dates), so we were confronted with a medium-sized vocabulary dictation task.

3 The user interface

Our GUI was really designed with the goal of serving many users on the same computer. The other main design aspect was to combine simplicity with good functionality. With our program only a microphone and a text editor (Microsoft Word or a similar word processing program) are needed for dictating medical reports.

Every user has one or more profiles containing all the special information characterizing his or her voice for a given language and vocabulary. The language models and the acoustic core modules can be installed separately, and the system can optionally adapt to the individual characteristics of the users. The user interface basically consists of a toolbar at the top of the desktop. Using the toolbar all the main functionalities related to the initial parameter settings can be accessed, such as choosing a specific user, choosing the actual task and selecting the output window (Fig. 1). Other functionalities can only be accessed from the actual text editor. The most important of these features could be that the user can ask the speech recognition system for other possible variants of the recognized sentences in cases where he/she discovers the recognized word or sentence to be incorrect.

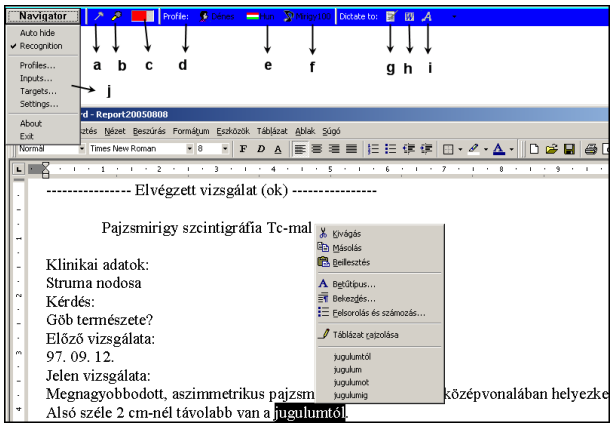

Figure 1: *Functions of the graphical user interface: a) Enable or disable auto hiding of the main toolbar. b) Start or stop the recognition procedure. The user can suspend the dictation at any time, and can continue later. c) Volume display bar. The volume of the microphone input can be checked here. d) Choosing a specific user. The user can be selected from the list of existing users. e) Choosing the actual language. The language assigned to the current user can be chosen from a listbox. f) Choosing the actual grammar. Any available grammar can be chosen with just one click. g) Selecting the internal text editor. The recognized text will appear in the 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. The functions presented above can be accessed from here.*

4 Acoustic modelling I: HMM phone models over MFCC features

At the level of acoustic modelling we have been experimenting with two quite different technologies. One of these is a quite conventional Hidden Markov Model (HMM) decoder that works with the usual mel-frequency cepstral coefficient (MFCC) features [4]. More precisely, 13 coefficients are extracted from 25 msec frames, along with their Δ and $\Delta\Delta$ values, at a rate of 100 frames/sec. The phone models applied have the usual 3-state left-to-right topology. Hungarian has the special property that almost all phones have a short and a long counterpart, and their difference is phonologically relevant (i. e. there are word pairs that differ only in the duration of one phone – for example 'tör'–'tőr' or 'szál'–'száll') [14]. However, it is known that such minimal word pairs are relatively rare [14], and inspecting the vocabulary of our specific dictation task we found no such words. Hence 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 modelling and this far 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 [4]. For speed efficiency it contains several built-in pruning criteria. First, it applies beam pruning, so only the hypotheses with a score no worse than the best score minus a threshold are kept. Second, the number of hypotheses extended at every time point is limited, corresponding to multi-stack decoding with a stack size constraint. The maximal evaluated phone duration can also be fixed. With the proper choice of these parameters the decoder on a typical PC runs faster than real-time on the medical dictation task.

5 Acoustic modelling II: HMM/ANN phone models over 2D-cepstrum features

Our alternative, more experimental acoustic model employs the HMM/ANN hybrid technology [2]. The basic difference between this and the standard HMM scheme is that here the emission probabilities are modelled by Artificial Neural Networks (ANNs) instead of the conventional Gaussian mixtures (GMM). In the simplest configuration one can train the neural net over the usual 39 MFCC coefficients – whose result can serve as a baseline for comparison with the conventional HMM. However, ANNs seem to be more capable of modelling the observation context than the GMM technology, so the hybrid models are usually trained over longer time windows. The easiest way of doing this is to specify a couple of neighboring frames as input to the net: in a typical arrangement 4 neighboring frames are used on both sides of the actual frame [2]. Another option is to apply some kind of transformation on the data block of several neighboring frames. Knowing that the modulation components play an important role in human speech perception, performing a frequency analysis over the feature trajectories seems reasonable. When this analysis is applied to the cepstral coefficients, the resulting feature set is usually referred to as the 2D-cepstrum [6]. Research shows that most of the linguistic information is in the modulation frequency components between 1 and 16 Hz, especially between 2 and 10 Hz. This means that not all of the components of a frequency analysis have to be retained, and so the 2D-cepstrum offers a compact representation of a longer temporal context.

In the experiments we tried to find the smallest feature set that would give the best recognition results. Running the whole recognition test for each parameter setting would have required too much time so, as a quick indicator of the efficiency of a feature set we used the frame-level classification score. Hence the values given in the following tables are frame-level accuracy values measured on a held-out data set of 20% of the training data.

First of all we tried to extend the data of the 'target' frame by neighboring frames, without applying any transformation. The results shown in Table 1 indicate that training on more than 5 neighboring frames significantly increased the number of features and hidden neurons (and also significantly raised the training time) without bringing any real improvement in the score.

Obs. size	Hidden neurons	Frame accuracy
1 frames	150	64.16%
3 frames	200	67.51%
5 frames	250	68.67%
7 frames	300	68.81%
9 frames	350	68.76%

Table 1: The effect of varying the observation context size.

In the experiments with the 2D-cepstrum we first tried to find the optimal size of the temporal window. Hence we varied the size of the DFT analysis between 8, 16, 32, and 64, always keeping the first and second components[1] (both the real and the imaginary parts), and combined these with the static MFCC coefficients. The results displayed in Table 2 indicate that the optimum is somewhere between 16 and 32 (corresponding to 160 and 320 milliseconds). This is smaller than the 400 ms value found optimal in [6] and the 310 ms value reported in [13], but this might depend on the amount of training data available (a larger database would cover more of the possible variations and hence would allow a larger window size). Of course, one could also experiment with combining various window sizes as Kanedera did [6], but we did not run such multi-resolution tests.

DFT size	Hidden neurons	Frame accuracy
8	200	64.63%
16	200	67.60%
32	200	67.01%
64	200	64.75%

Table 2: Frame-level results at various DFT sizes.

As the next step we examined whether it was worth retaining more components. In the case of the 16-point DFT we kept 3 components, while for the 32-point DFT we tried retaining 5 components (the highest center frequency being 18.75 Hz and 15.625 Hz, respectively). The results (see Table 3) show that the higher modulation frequency components are less useful, which accords with what is known about the importance of the various modulation frequencies.

Finally, we tried varying the type of transformation applied. Motlíček reported that there is no need to keep both the real and imaginary parts of the DFT coefficients; using just one of them is sufficient. Also, he obtained a similar performance when replacing the complex DFT with the DCT [10]. Our findings agree more with those of Kanedera [6], that is we obtained slightly worse results with these modifications (see Table 4). Hence we opted for the complex DFT, using both the real and imaginary coefficients. One advantage of the complex DFT over the DCT might be that when only some of its coefficients are required (as in our case), it can be very efficiently computed using a recursive formula [5].

[1]The DC offset being indexed as the zeroth component.

DFT Size	Components	H. n.	Frame acc.
16	1, 2, 3	250	68.40%
32	1, 2, 3, 4, 5	300	70.64%

Table 3: Frame-level results with more DFT components.

Transform	H. neurons	Frame accuracy
DFT Re + Im	300	70.64%
DFT Re only	220	65.81%
DCT	220	68.00%

Table 4: The effect of varying the transformation type.

6 Domain-specific language modelling

A special difficulty of creating language models for Hungarian is the highly agglutinative [3] nature of the language. This means that most words are formed by joining several morphemes together, and those modifications of the meaning that other languages express e.g. by pronouns or prepositions in Hungarian are handled by affixes (for example 'in my house' is 'házamban') [7]. Because of this, in a large vocabulary modelling task the application of a morphologic analyzer/generator seems inevitable. First, simply listing and storing all the possible word forms would be almost impossible (e.g. 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 [11]. Hence the training of conventional n-gram 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 the morphology would be necessary.

Though decent morphological tools exist now for Hungarian, in our medical dictation system we preferred to avoid the complications incurred by morphology. In fact, the restricted vocabulary is one of the reasons why we opted for the medical dictation task. For, as we men-

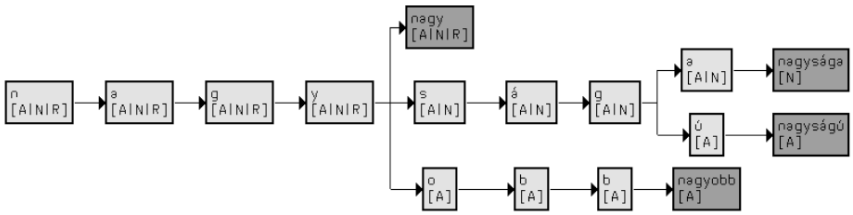

Figure 2: *Prefix tree for some Hungarian words with their MSD code. At the branches of the tree the grammar model can generate the probability of a word based on the word n-gram and also based on the class n-gram.*

tioned earlier, the thyroid gland medical reports contain only about 2500 different word forms. Although these many words could be easily managed even by a simple list ('linear lexicon'), we organized the words into a lexical tree where the common prefixes of the lexical entries are shared. Apart from storage reduction advantages, this representation also speeds up decoding, as it eliminates redundant acoustic evaluations [4]. A prefix tree representation is probably far 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 encountering 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 worst case to a small ϵ constant) when necessary. During the evaluation of the n-grams the system applies a language model lookahead technique. This means that the language model returns its scores as early as possible, not just at word endings. For this reason the lexical trees are stored in a factored form, so that when several words share a common prefix, the maximum of their probabilities is associated with that prefix [4]. These techniques allow a more efficient pruning of the search space.

Besides word n-grams we also experimented with constructing class n-grams. For this purpose 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 [8]. This software associates one or more MSD (morpho-syntactic description) code with the words, and we constructed the class n-grams over these codes. 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 [1]. Figure 2 shows an example of a prefix tree storing four words, along with their MSD codes.

7 Experimental results and discussion

For testing purposes we recorded 20-20 medical reports from 2 male and 2 female speakers. The language model applied in the tests was constructed based on just 500 reports instead of all the 8546 we had collected. This subset contained almost all the sentence types that occur in the reports, so this restriction mostly reduced the dictionary by removing a lot of rarely occurring words (e.g. dates and disease names). Besides the HMM decoder we tested the HMM/ANN hybrid system in three configurations: the net being trained on one frame of data, on five neighboring frames, and on the best 2D-cepstrum feature set (static MFFC features plus 5 modulation components using a 32-point DFT with both *Re* and *Im* parts). The results are listed in Table 5 below. Comparing the first two lines, we see that when using the classic MFCC features the HMM and the HMM/ANN system performed quite similarly on the male speakers. For some reason, however, the HMM system did not like the set of female voices. The remaining rows of the table show that extending the net's input with an observation context – either by neighboring frames or by modulation features – brought only very modest improvements over the baseline results. We think the reason for this is that in the current arrangement the recognizer relies very strongly on the language model, thanks to the high predictability of the sentences. We suspect that the improvement in the acoustic modelling will be better seen in the scores when we apply the system to a linguistically less restricted domain. Pure phone recognition tests (i.e. recognition experiments with no language model support) that could verify this conjecture are just under development.

8 Conclusions

This paper reported the current state of a Hungarian project for the automated dictation of medical reports. We described the acoustic and linguistic training data collected and the current state of development in both the acoustic and linguistic modelling areas. Recognition results were

Model Type	Feature Set	Male 1	Male 2	Female 1	Female 2
HMM	MFCC + $\Delta + \Delta\Delta$	97.75%	98.22%	93.40%	93.39%
HMM/ANN	MFCC + $\Delta + \Delta\Delta$	97.65%	97.37%	96.78%	96.91%
HMM/ANN	5-frames $*$ (MFCC + $\Delta + \Delta\Delta$)	97.65%	97.74%	96.67%	98.05%
HMM/ANN	MFCC + 5 Mod. Comp. (Re + Im)	97.88%	97.83%	96.86%	96.42%

Table 5: Word recognition accuracies of the various models and feature sets.

also given over a somewhat restricted subset of the full domain. For the next step we plan to extend the vocabulary and language model to cover all the available data, and then to test the system over other dictation domains as well. Our preliminary results indicate that for tasks over larger vocabularies several further improvements will be required. On the acoustic modelling side we intend to implement speaker adaptation and context-dependent models within the HMM system. We also plan to continue our research on observation context modelling within the HMM/ANN system. Finally, the language model will also need to be improved in many respects, especially when handling certain special features like dates and abbreviations.

References

[1] A. Bánhalmi, A. Kocsor, and D. Paczolay. 2005. Supporting a Hungarian dictation system with novel language models (in Hungarian). In: *Proc. of the 3rd Hungarian Conf. on Computational Linguistics*, pp. 337–347.

[2] H. Bourlard and N. Morgan. 1994. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic.

[3] D. Crystal. 2003. *A Dictionary of Linguistics and Phonetics.* Blackwell Publishing.

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

[5] E. Jacobsen and R. Lyons. 2004. An update to the sliding DFT. *IEEE Signal Processing Magazine*, 21(1):110–111.

[6] N. Kanedera, H. Hermansky, and T. Arai. 1998. Desired characteristics of modulation spectrum for robust automatic speech recognition. In: *Proc. ICASSP'98*, pp. 613–616.

[7] A. Kornai. 1994. *On Hungarian morphology.* Hungarian Academy of Sciences.

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

[9] Medisoft. 2004. www.medisoftspeech.hu

[10] P. Motlíček. 2003. *Modeling of Spectra and Temporal Trajectories in Speech Processing*. Ph.D. Thesis, Brno University of Technology.

[11] G. Németh and Cs. Zainkó. 2001. Word unit based multilingual comparative analysis of text corpora. In: *Proc. Eurospeech 2001*, pp. 2035–2038.

[12] Nuance. 2007. http://www.nuance.co.uk/naturallyspeaking/

[13] P. Schwarz, P. Matějka, and J. Černocký. 2003. Recognition of phoneme strings using TRAP technique. In: *Proc. Eurospeech 2003*, pp. 825–828.

[14] P. Siptár, M. Törkenczy. 2000. *The phonology of Hungarian.* Oxford University Press.

[15] M. Szarvas and S. Furui. 2002. Finite-state transducer based Hungarian LVCSR with explicit modeling of phonological changes. In: *Proc. ICSLP 2002*, pp. 1297–1300.

[16] K. Vicsi, A. Kocsor, Cs. Teleki, and L. Tóth. 2004. Hungarian speech database for computer-using environments in offices (in Hungarian). In: *Proc. 2nd Hungarian Conf. on Computational Linguistics*, pp. 315–318.

[17] K. Vicsi, Sz. Velkei, Gy. Szaszák, G. Borostyán, G. Gordos. 2006. Speech recognizer for preparing medical reports – Development experiences of a Hungarian speaker independent continous speech recognizer. *Híradástechnika, Vol. 61, No. 7, pp. 22-27.*