First Results of a Hungarian Medical Dictation Project

András Bánhalmi, Dénes Paczolay, Lászó Tóth, 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

Abstract

This paper reviews the current state of a Hungarian project that seeks to create a speech recognition system for the dictation of thyroid gland medical reports. We present the MRBA speech corpus that was collected to support the training of Hungarian LVCSR systems. Besides the speech data, a huge set of medical reports was also collected to help the creation of domain-specific language models. At the acoustic modelling level we experiment with two techniques – a conventional HMM one and an ANN-based solution – which are both briefly described in the paper. Then we present the language modelling methodology currently applied in the system, and round off with recognition results on test data taken from four people. The scores show that on the current restricted domain we are able to produce word accuracies over 95%, but the planned extension of the system to larger vocabularies will probably require further improvements.

Prvi rezultati madžarskega projekta narekovanja zdravniških izvidov

Prispevek predstavlja pregled trenutnega stanja madžarskega projekta, ki skuša vzpostaviti sistem razpoznavanja govora za narekovanje zdravniških izvidov na temo žleze ščitnice. Predstavljamo govorni korpus MRBA, ki je bil sestavljen za podporo učenju madžarskih sistemov za razpoznavanje govora z velikim besednjakom. Poleg govornih podatkov je bilo zbrano tudi veliko število zdravniških izvidov za pomoč pri pripravi jezikovnih modelov za omenjeno področje uporabe. Na ravni akustičnega modeliranja eksperimentiramo z dvema tehnikama - konvencionalno s prikritimi Markovovimi modeli in rešitvijo, ki temelji na nevronskih mrežah - obe sta kratko predstavljeni. Nato predstavljamo metodologijo jezikovnega modeliranja, ki je trenutno uporabljena v sistemu, in zaključimo z rezultati razpoznavanja na testnih podatkih štirih govorcev. Rezultati pokažejo, da smo pri trenutnem omejenem področju uporabe zmožni dosegati točnost razpoznavanja besed višjo kot 95%, načrtovana razširitev sistema na širše besedišče pa bo verjetno zahtevala dodatne izboljšave.

1. Introduction: goals of the 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 only for restricted vocabularies (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 modelling due to the highly agglutinative nature of Hungarian.

In 2004 the Research Group on Artificial Intelligence, 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 lasts for three years, and is financially supported by the national fund IKTA-056/2003. As regards acoustic modelling, the project includes the collection and annotation of a large speech corpus of phonetically rich sentences. As regards language modelling, we restricted the target domain to the dictation of certain types of medical reports. Although this clearly leads 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 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 teams use the same speech corpus for training, they focus on different dictation tasks and experiment with their own acoustic and language modelling technologies. Our team (Szeged) deals with the dictation of thyroid scintigraphy medical reports, while the Budapest team deals with gastroenterology reports. This paper describes the current state of development of the Szeged team only.

2. Speech and language resources

In the first phase of the project we designed, collected and annotated a speech database that we refer to as the MRBA corpus (the abbreviation stands for the "Hungarian Reference Speech Database") (Vicsi et al., 2004). 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 work 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 are between 16-30 years in age, the rest being evenly distributed among the remaining age groups. Both home PCs and laptops were used for the recordings, and the microphones and the 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.

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 done here.

For the construction of the domain-specific language models, we obtained 9231 written medical reports from the Department of Nuclear Medicine of the University of Szeged. These thyroid scintigraphy reports were written and stored between 1998 and 2004 using various software packages that were employed at the department during that period. So first of all we had to convert all the reports to a common format, which was followed by several steps of 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, 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 privacy. Then we discarded those reports that were incomplete like those that had missing fields. This way only 8546 reports were kept, which contain 11 sentences and 6 words per sentence on average. The next step was to remove any typographical errors from the database, of which there were surprisingly many (some words occurred in 10-15 mistyped forms). A special problem was that of unifying those Latin terms that can be written both with a Latin or a Hungarian spelling. The abbreviations had to be resolved, too. The corpus we got after these steps contained approximately 2500 different word forms, so we were confronted with a medium-sized vocabulary dictation task.

3. 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 over the usual mel-frequency cepstral coefficient (MFCC) features (Huang et al., 2001). 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. Although Hungarian has the special property that almost all phones have a short and a long counterpart, in the vocabulary of our specific dictation task they seemed to have no discriminative role. 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 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. 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 limited. Normally the decoder runs faster than real-time on our dictation task on a typical PC.

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

Our alternative, more experimental acoustic model employs the HMM/ANN hybrid technology (Bourlard and Morgan, 1994). 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. 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 solution for this is to specify a couple of neighboring feature frames as input to the net: a conventional arrangement is to use 4 neighboring frames on both sides of the actual frame (Bourlard and Morgan, 1994). 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 (Kanedera et al., 1998). Research shows that most of the useful 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 gave the best recognition results. As a quick indicator of the efficiency of a representation we used the frame-level classification score, so the values given below 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 only significantly increased the number of features

and hidden neurons (and even more considerably the training time) without bringing a real improvement in the score.

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 retaining the first and second components (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 must be somewhere between 16 and 32 (160 and 320 milliseconds). This is smaller than the 400 ms value found optimal by Kanedera et al. (1998) and the 310 ms value reported by Schwarz et al. (2003), 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 the combination of various window sizes as Kanedera et al. (1998) did, but we did not run such multi-resolution tests.

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 show (Table 3) 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 retain 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 DCT (Motlíček, 2003). Our findings agree more with those of Kanedera et al. (1998), that is we obtained slightly worse results with these modifications (see Table 4). So 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 formulation (Jacobsen and Lyons, 2004).

Obs. size	Hidden neurons	Frames correct
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.

DFT size	Hidden neurons	Frames correct
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.

DFT Size	Components	H. neurons	Frames corr.
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	Hidden neurons	Frames correct
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.

5. Domain-specific language modelling

A special difficulty of creating language models for Hungarian is the highly agglutinative nature of the language. In a large vocabulary modelling task the application of a morphologic analyzer/generator seems inevitable. First, simply listing and storing all the 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). 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 morphology would be necessary.

Though quite good morphological tools exist now for Hungarian, in the first experiments with our 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 was mentioned, the thyroid gland medical reports contained only about 2500 different word forms. Although these many words could be easily managed even by a simple list ('linear lexicon'), we organize them 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 (Huang et al., 2001). The prefix tree representation is quite 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 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 purpose the lexical trees get factored, so that when several words share a common prefix, the maximum of their probabilities is associated with that prefix (Huang et al., 2001). These tech-

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.

niques 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 (Kuba et al., 2004). 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 (Bánhalmi et al., 2005).

6. Experimental results and discussion

For testing purposes we recorded 20-20 reports from 2-2 male and female speakers. The language model applied in the tests was constructed based on only 500 reports instead of all the 8546 ones we 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 complex DFT, both Re and Im parts). The results are listed in Table 5. Comparing the first two lines, we see that when using the same 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. Extending the net's input with an observation context – either by neighboring frames or by modulation features – brought only a modest improvement over the baseline results. We think that the improvement in the acoustic modelling will be more prominently reflected in the scores when moving to a linguistically less restricted domain where the decoder cannot rely so strongly on the language model as it does in the current configuration.

7. 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. Preliminary recognition results were also given over a somewhat restricted subset of the full domain to be handled. As the next step we plan to extend the vocabulary and language model to cover all the available data, and our preliminary results show that for a larger vocabulary several further improvements will be necessary. 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, especially when handling certain special features like dates or abbreviations.

8. References

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.

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

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

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

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

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

Medisoft. 2004. www.medisoftspeech.hu.

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

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

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

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

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