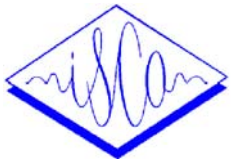

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ANALYSIS OF HMM MODELS IN ALPHABET LETTERS RECOGNITION

Stefan Grocholewski
Institute of Computing Science, Poznan University of Technology
Piotrowo 3a, 60-965 Poznan, Poland
grocholew@put.poznan.pl

ABSTRACT

In the paper we consider the problems concerning the speaker independent alphabet letters recognition by using the common CDHMMs along with standard cepstral feature vectors. The propositions of new phonetically motivated features are discussed.

Keywords: alphabet letters recognition, HMM's, cepstral feature vectors

1. INTRODUCTION

At the previous Eurospeech'97 we presented the first fully annotated speech database for Polish [1]. As the signal representation is a crucial issue in the design of speech recognizers, our first experiments concerned the appropriate front-end for future speech recognition systems. This is very important task because all the other recognition steps depend on the quality of the feature extraction level in the signal processing stage. Currently DFT / filter bank derived mel-based cepstral coefficients [2] are probably the most common form of the front-end. The important property of cepstral analysis is that a multiplication in the power spectral domain becomes an addition in the cepstral domain; it helps to remove the influence of the transmission channel on the speech signal. To obtain the best recognition performance, typically, derivatives of cepstral coefficients [3] are concatenated to the static cepstral coefficients to form the speech parametric representation.
Among many techniques based on cepstral parameters we can mention: cepstral liftering [4], cepstral domain filtering [5], Cepstral Mean Normalization and other cepstral normalization / compensation techniques, linear discriminant analysis to optimize the speech feature set, two dimensional cepstrum, etc.
In our paper we try to test cepstral parameters based front-end performance in the task of speaker independent Polish alphabet letters recognition.
In the next section we describe shortly our database CORPORA used in experiments.
In the section 3 we describe the reference "auditory" system and in the following section we compare this reference recognizer with the very common CDHMM based system. At the end we present our modifications connected with the front-end.

2. USING CORPORA

As we have mentioned in the first section, we have used the first speech database for Polish to create the learning set (693 utterances - 21 speakers including men, women and children, 33 letters) and the testing set (660 utterances -20 speakers). The utterances have been recorded in the computer room in various laboratories in Poland with the sampling frequency 16 kHz and 16 bit resolution. The details about how to use CORPORA for comparing ASR systems can be find in [6].

3. HUMAN LISTENERS AS THE REFERENCE RECOGNIZER

A group of 11 listeners performed as a reference recognizer. They had to recognize, in the auditory experiment, 20 realizations of each of 33 letters, i.e. 660 recordings from the testing set.
The letters were presented in the stochastic order. The results of the recognition session are presented in the table 1.

4. DESCRIPTION OF THE BASELINE SYSTEM

As the baseline system we have used Continuous Density HMM's (HTK 2.1 software toolkit [7]). In the first experiment we tried the context independent monophones based system. The feature vector consisted of energy, 12 cepstral

parameters and their first and second derivatives and the models consisted of 5 states / 1 Gaussian mixture pro state.
The results of the speaker independent recognition are presented in the Table 2

Letters	Recognition error
em	22,7 %
e~	10%
ɨ	8,5%
en	8%
ef	7%
pe, te, er, eɕ	6%
o	4,5%
i, eʃ	2,5%
el	2%
de, es	1,5%
ew, eŋ	1%
be, u	0,5%
all other	0%
Mean error	**2,95**

Table 1. Recognition errors in the auditory experiment

Letters	Recognition error
te	100%
pe	80%
o	45%
de	35%
en, eŋ	30%
Er	25%
be, el	20%
e~, ka, ɨ	15%
a~, e, eʃ,	10%
a, ge, em, ku, es, ʒet	5%
all others	0%
Mean error	**16,7%**

Table 2. Recognition errors for the baseline system

As we can see the simplest system gave, in general, much poorer results comparing with the auditory system of humans. Nevertheless it is worthwhile to mention some exceptions; the set {em, *ef*, ew, *eɕ*, *u*} was better recognized comparing with human listener.

The plosives *pe*, *te* were recognized very badly because the only one HMM model existed for the vowel *e*. To explore the information about the place of articulation existing in the starting part of *e* in the words *pe*, *te*, we calculated the models of triphones, especially for the vowel *e*.
In the next step we experimented with very common triphone based CDHMM models with 3 Gaussian mixtures pro state. It resulted in reduction in *p* and *t* missrecognition to about 1/3.
Generally in the case of triphone base models, calculated for 21 speakers, with three emitting states and three Gaussian mixtures the recognition error was 10,9%.
The question is - this relatively high recognition error is due to small training set, or this small training set is sufficient, but requires more sophisticated front-end, or maybe the learning process is not perfect.
To evaluate the learning process we have used the same learning set as the testing set. We considered two learning sets: for 21 and for 45 speakers.
The results are presented in the Table 3.

	21 speakers	45 speakers
Monophones	4,8%	7,5%
Triphones	0,14%	1,1%

Table 3. The recognition errors in the case when the learning and testing sets were the same

Since our database consists of recordings from 45 speakers, we repeated 45 times the learning and recognition steps using cross validation method, i.e. each time the other speaker constituted the testing set whereas all other constituted the learning set. In this case 45 speakers were recognized using the learning set with 44 speakers. It enabled a comparison of two learning sets. The results are presented in the Table 4.

	21 speakers (ls) 20 speakers (ts)	44 speakers (ls) 45 speakers (ts)
Monophones	16,7%	13,5%
Triphones	10,9%	8,8%

Table 4. The recognition errors in the case when the learning and testing sets were different; ls, ts – signify the learning and testing sets

The best of above results (8,8% recognition error in speaker independent alphabet letters recognition) shows that it is much worse

comparing with the "auditory" system – see Table 1.
In the next section we present some modifications in the front-end .

5. IMPROVEMENTS IN THE BASELINE SYSTEM

The main improvements are concerned with the most frequent errors in two groups of letters: {pe, te, e}, {em, en, el}.
In the Fig. 1 - 4 we show the amplitude normalized mean spectra of some phonemes obtained from the mean static cepstral coefficients of their models, by summing the cosinusoids corresponding to c1 – c12. Note that c0 is not used in our case and that all cepstral coefficient are liftered – the influence of higher coefficient is clearly visible. Since we use MFCC the spectra are on the mel scale. The highest frequency is 8 kHz.
There are nine spectra pro phoneme; they correspond to three states of the model (x axis) and three Gaussian mixtures (y axis).
In the Fig. 1 we show the superposition of the appropriate cepstra for very distinguishable vowels *a* and *e*. The visual inspection confirms this ability to distinguish both vowels.
In the Fig. 2 we show the opposite case of very similar mean static spectra of triphones *t-e* and *p-e*. There is a very little difference between *e* in the first states of both models (see section 4).
In the Fig. 3 and Fig. 4 we show the appropriate spectra for other missrecognized pairs: {*m,n*} and {*n,l*}.
The visual inspection of these examples suggests that generally the Euclidean distance between cepstral vectors is suitable for recognition purposes, but there are the cases when the small range of frequency is discriminative and should be analyzed. The cepstral coefficients, as calculated equally over all frequencies of the log spectrum, are not the best solution in such a case. This is due to the fact that there is no simple relation between formants or local features and particular cepstral coefficients.
The aim of our first efforts has been the decrease of recognition errors of stops and nasals.
During the experiments we introduced into the feature vector some phonetically motivated features obtained by visual inspection of mean static spectra for all Gaussian mixtures. For the nasals we noticed the reduction of about 40 percent of errors.
As we noticed in section 3, Polish stops are highly missrecognized. To decrease the recognition errors we increased the role of bursts in stops recognition. According to [8] the first few ms of burst provide most of the information about the place of articulation. Because the time resolution in our experiments was insufficient to explore this fact (10 ms), we have manually prolonged the beginning parts of bursts to better model the stops. Although it requires the off-line recognition it resulted in decreasing the number of errors in stops recognition of about 25%.
After all modifications mentioned above we obtained 6,4% of errors in speaker independent alphabet letters recognition for 45 speakers using cross validation method.

6. CONCLUSIONS

Alphabet letter recognition in speaker independent recognition system, even for the clean speech is a challenge. The appropriate system has to be able to perform fine phonetic distinctions. In the paper we described our experiments with such a system. We compared the common HMM's with the reference "auditory" system. The results showed that in some cases machines can even outperform humans. To deal with the most difficult phonemes we enlarged the standard cepstral feature vector with some phonetically motivated features.
We plan to elaborate and study another features related to dynamic cepstra.

ACKNOWLEDGMENT

This work was supported by KBN Grant No 8 T11E 042 15

REFERENCES

[1] S.Grocholewski (1997), Corpora – Speech Database for Polish Diphones, *Proc. EUROSPEECH'97*, pp. 1735 – 1738.
[2] S.Davis, P.Mermelstein (1980), Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences, *IEEE Trans. on ASSP*, vol.28, 1980, no 4, pp.357 – 366.
[3] S.Furui (1986), On the role of spectral transition for speech perception, *J.Acoust.Soc.Am.,* vol.80, 1986,no 4, pp1016-1025.

[4] B.H.Juang, L.Rabiner, J.Wilpon (1986), On the use of bandpass liftering in speech recognition. *Proc. ICASSP'86*, pp. 765 - 768.
[5] B.Hanson, T.Applebaum (1993), Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech. *Proc. ICASSP'93*, pp. II.79-II.82
[6] S.Grocholewski (1999), The use of CORPORA for comparing ASR systems*, Speech and Language Technology (ed. W.Jassem at all),* in printing.
[7] S.Young (1997), The HTK Book, Cambridge Research Laboratory.
[8] E.Łukasik, S.Grocholewski (1998), Comparison of Some Time - Frequency Analysis Methods for Classification of Plosives*, Proc. Eusipco'98,* pp. 709-712.

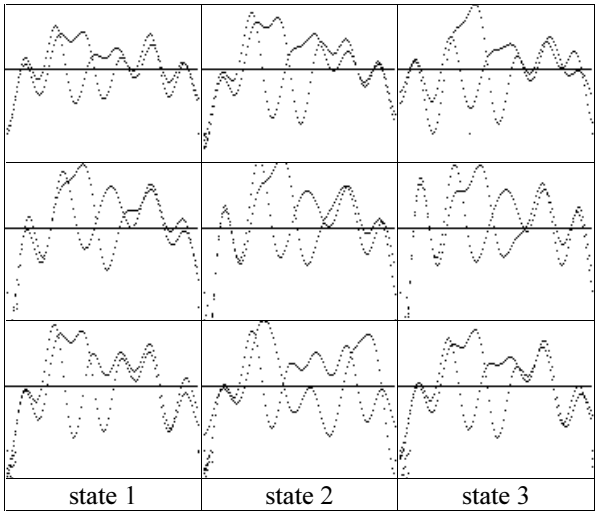

Fig. 1. The means spectra on the mel scale of the three Gaussian mixtures (y axis) in the 3 state (x axis) models of a and e

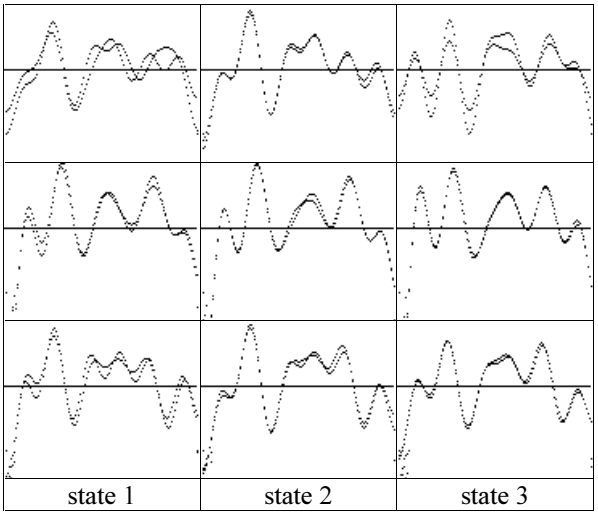

Fig. 2. The means spectra on the mel scale of the three Gaussian mixtures (y axis) in the 3 state (x axis) models of t-e and p-e

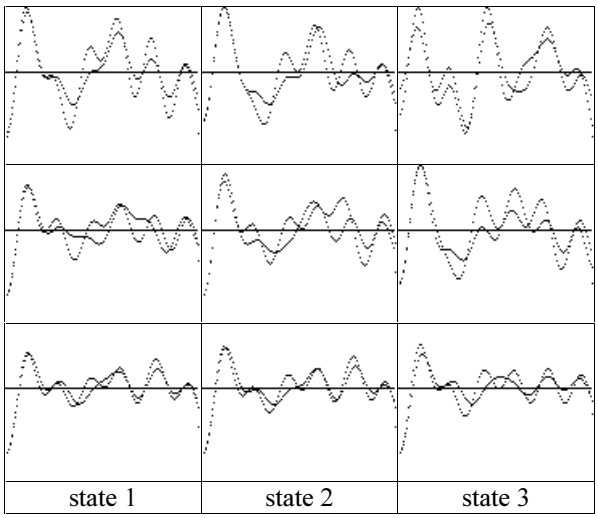

Fig. 3. The means spectra on the mel scale of the three Gaussian mixtures (y axis) in the 3 state (x axis) models of e-m and e-n

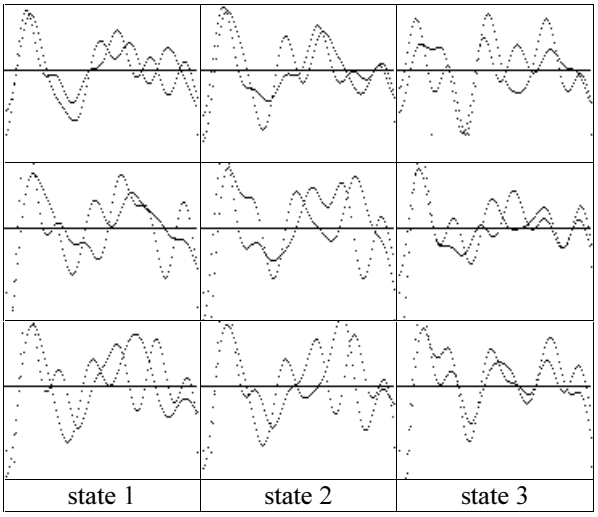

Fig. 4. The means spectra on the mel scale of the three Gaussian mixtures (y axis) in the 3 state (x axis) models of e-n and e-l