https://doi.org/10.26552/com.C.2004.4.17-21

Roman Jarina – Michal Kuba *

SPEECH RECOGNITION USING HIDDEN MARKOV MODEL WITH LOW REDUNDANCY IN THE OBSERVATION SPACE

Current speech recognition systems usually model a speech signal as a finite-state stochastic process, in which acoustic observations are obtained through short-term spectral analysis. The model has to deal with several thousands of speech parameters during one second of utterance. A great redundancy in the parameters makes processing computationally very expensive. We propose a combination of 2-D cepstral analysis and continuous Hidden Markov Model with a small, optimally designed, number of states and acoustic observations. 2-D cepstrum efficiently preserves spectral variations of speech and yields uncorrelated parameters in both time and frequency. The system is evaluated on isolated word recognition task in Slovak language. Promising preliminary results are presented.

1. Introduction

A majority of the state-of-art ASR systems models a speech signal as a finite-state stochastic process to handle the great variability found in human speech. Acoustic observations of speech are obtained through short-term spectral analysis. One speech feature vector, which forms one observation, usually consists of static spectral features (e.g. 13 cepstral coefficients) and their time derivatives that determine temporal variations of speech spectra. Such parameterisation is not optimal from either statistical or perceptual point of view. A great redundancy in the speech features makes further processing computationally much more costly.

Two-dimensional cepstral analysis preserves spectral variations more efficiently while also yields uncorrelated features in both time and frequency. In this paper we are giving ourselves the question whether speech signal can be "observed in time" by a much smaller number of the feature vectors (or observations) than it is common in the present ASR systems. We propose a combination of 2-D cepstral analysis and left-right continuous Hidden Markov Model with a small (optimally designed) number of states and acoustic observations. The system is evaluated on an isolated word recognition task in Slovak language. Promising preliminary results are presented.

2. On Markov Modelling of Speech

The most commonly used model for ASR is the first-order Markov process. The popularity of this method lies in its model simplicity, ease of training and acceptable recognition precision on certain tasks [1]. Thus time-varying characteristics of a speech signal are described through a chain of static states. Each model has a number of states that approximates the number of distinct acoustic or phonetic events in the unit being modelled. Such units are commonly words or subword units as phonemes, diphones, etc. Since details of the Markov model's operation in speech analysis must be inferred through observations of speech, the states of the model are hidden. Such model is usually referred to as Hidden Markov Model (HMM). A HMM constitutes of the state-observations and transition probabilities. The transition probabilities provide a mechanism for connection of the states, and for modelling variations in speech duration and articulation rates. The statistical distributions of speech features define acoustic observations. Mel-frequency cepstral coefficients (MFCC) are the most popular features of speech [1] [2]. The following discussion deals with the HMM with continuous distributions of speech features.

The continuous HMM is defined by a set of parameters as follows

$$\lambda = (A, B, \pi, N), \tag{1}$$

where A is the transition probabilities matrix, B is the output probabilities matrix, π is the vector of initial probabilities, and N is the total number of states (Figure 1). Let $\{y_1, y_2, ..., y_t\}$ and $\{s_1, s_2, ..., s_t\}$ be time sequences of acoustic observations and related hidden states respectively then the probability of taking a transi-

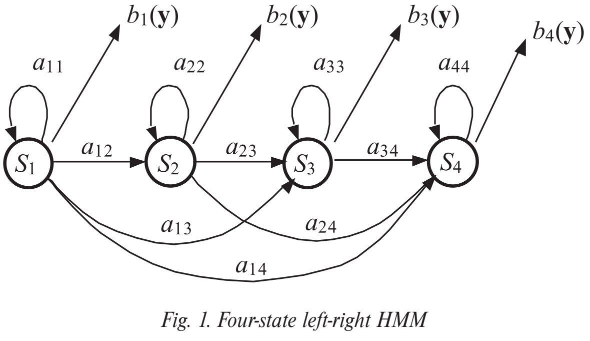

Fig. 1. Four-state left-right HMM

* **Roman Jarina, Michal Kuba**
Department of Telecommunications, Faculty of Electrical Engineering, University of Žilina, Žilina, Slovakia
E-mail: roman.jarina@fel.utc.sk, michal.kuba@fel.utc.sk

tion from the state i to the state j is $a_{ij} = P(s_t = j \mid s_t - 1 = i)$, and $A = \{a_{ij}\}$; The output probability of emitting the feature vector y_t when state i is entered, is $b_i(y_t) = P(y_t \mid s_t = i)$, and $B = \{b_i(y_t)\}$.

In the first-order MM is the assumption $P(s_t \mid s_{t-1}, s_{t-2}, ..., s_1) = P(s_t \mid s_{t-1})$. That means only neighbouring states depend on each other, and past history, except the neighbour previous state, is ignored in signal modelling. Although this simplification enables much easier computation, it represents rather inaccurate modelling since speech perception is conditional on much longer time period. Human short-memory lasts several seconds whereas ASR HMM "views" only the past speech frames. To enable to incorporate a longer past time period (2 or more previous frames), 1^{st} and 2^{nd} time derivates (or differences) of MFCC, referred to as *delta* (Δ) and *delta delta* (Δ^2) coefficients, are added to the speech feature set. Thus speech signal is represented by time sequence of the feature vectors. Each feature vector usually consists of 13 MFCC, which represent short-time spectrum, 13 ΔMFCC and 13 Δ^2MFCC, which represent spectral dynamics. The feature vectors are computed on a frame-by-frame basis. In such a case the one second long utterance is represented by almost 4,000 parameters (if the frame length is 10 ms).

3. Spectral Dynamics Represented by the Modulation Spectrum

Obviously, the prime carrier of the linguistic information is changes of the vocal tract shape. Such changes are reflected in changes of the spectral envelope of the speech signal. Furui has already shown that spectral transitions play very important role in speech perception [3]. If the spectral envelope is represented by a set of coefficients (e.g. MFCC or Filter-Bank energies), each coefficient varies gradually within each distinct segment of speech and thus forms the time contour (magnitude of the coefficient as a function of time). A shape of such contours, or spectral dynamics, is usually described explicitly by adding delta coefficients to the feature vector.

The procedure of delta coefficients computation can be seen as a simple FIR filtering applied on time trajectory of each of the spectral component (FB-energies in spectral domain or MFCC in cepstral domain). More general approach to filtering of these time trajectories, known as RASTA processing, was introduced and extensively studied by Hermansky et al. [4], [2]. For illustration, a spectrogram of the Slovak word "osem" is shown in Figure 1. Spectral transitions between vowels are clearly visible.

3.1 Two-Dimensional Cepstrum

The mel-frequency 2-D cepstral coefficients are computed by applying 2-D cosine transform on the block of consecutive spectral vectors (mel-FB energies) as follows [5]

$$\hat{S}_{FB}(k, m) = \log(|S_{FB}(k, m)|), 0 \le k \le K-1,$$
$$0 \le m \le L-1, \tag{2}$$

where $S_{FB}(k, m)$ is the mel-spaced filter bank (FB) spectrum of the frame m, K is the number of critical-width bands and L is the number of frames used in the analysis block

$$c(u, m) = \frac{1}{K}\sum_{k=0}^{K-1} \hat{S}_{FB}(k, m) \cdot \cos\left[\frac{(2k+1)\pi u}{2K}\right],$$
$$0 \le u \le K-1, 0 \le m \le L-1, \tag{3}$$

$$C(u, m) = \frac{1}{L}\sum_{m=0}^{L-1} c(u, m) \cdot \cos\left[\frac{(2m+1)\pi v}{2L}\right],$$
$$0 \le u \le K-1, 0 \le v \le L-1, \tag{4}$$

Since not all the coefficients of the matrix $C = \{C(u,v)\}$ are needed for ASR, only selected coefficients form the *TDC (Two-Dimensional Cepstrum)* feature vector.

Spectral analysis of temporal trajectories of spectral envelopes yields the *modulation spectrum* of speech. In a TDC matrix, the dimension v (in Eq. 4) represents the modulation spectrum. Between the index v and modulation frequency in Hz is the following equation

$$\theta = \frac{F_s}{n_F} \cdot N \cdot v = \frac{v}{T}, \tag{5}$$

where θ is the modulation frequency in Hz, F_s is sampling rate, n_F is the number of frames in the analysis block, N is the length of the frame, T is the total duration of the analysis block in seconds. The human auditory system is most sensitive to modulation frequencies around 4 Hz that reflects the syllabic rate of speech. Thus the human hearing in perception of modulated signals acts as a band-pass filter with the length of the impulse response of minimally 150-250 ms [2] which is the length of 15-20 frames. The results of speech recognition experiments have shown that the components of the modulation spectrum below 1 Hz and above 16 Hz have only a minor role in both human perception and ASR [6].

In [5], [7], we have also studied discriminative properties of TDC features on discrimination of confusable Slovak consonants. In test utterances a group of 6 consonants (3 stops and 3 fricatives) were placed between the same two vowels. We studied what components of the TDC matrix are the most important for a phoneme discrimination. We used 12 cepstral coefficients along the frequency axis (dimension u in Eq. 3–4). We confirmed that only coefficients corresponding to the modulation frequencies from the range mentioned above are important. In our case the sufficient subset of TDC coefficients was between 12×5 and 12×7 whereas the coefficients with the index $v = 0$ (i.e. $\theta = 0$ Hz) were excluded. If the coefficients with the index $v = 0$ were included into the set, the recognition rate decreased rapidly, particularly for noisy speech.

3.2 2-D Cepstrum analysis in HMM framework

For ASR task, 2-D cepstrum (TDC) was first introduced by Ariki [8] who used only one TDC matrix for each word. Recently, Jarina [5] made several experiments in which he modelled words by a small number (1 or 3) of linearly spaced TDC matrices. He applied a multilayer perceptron to discriminate the patterns formed from these matrices. Also several works on combination of TDC and continuous HMM have been reported [9–11]. Milner [9] used the TDC for each frame while Kanedera [10] used a much longer temporal window and only a small selection of TDC, which corresponds with the range of modulation frequencies between 3 and 9.5 Hz. The authors of these experiments reported increase of recognition rate when using TDC based dynamic features rather than conventional delta features.

In this paper, we investigate the combination of 2-D cepstrum and HMM from a different point of view. HMM assumes that the acoustic observations are uncorrelated. But in reality, an intra-frame correlation (i.e. between static and dynamic features) as well as a high inter-frame correlation (i.e. between successive frames) are observed. Thus the number of observations in conventional ASR systems is over-estimated. There are about 50–100 observations for one second utterance.

TDC analysis enables modelling dynamic properties of the signal implicitly. The TDC computed via 2-D cosine transform produces almost uncorrelated set of coefficients in both frequency (index u) and modulation frequency (index v) dimensions. We hypothesise that if the TDC is applied, a much smaller number of observations is necessary (due to decorrelation properties of TDC, a high redundancy in temporal trajectories of the speech envelopes can be removed). For instance, in [5] we have shown that 22 spectral vectors could be replaced by only one TDC matrix, from which only about 60–70 coefficients are needed, unlike conventional ASR systems, in which almost 800 coefficients ($MFCC+\Delta+\Delta^2$) are required for the same duration of speech. Each observation, which is formed from TDC features, will incorporate information from several hundreds milliseconds of speech which is in accordance with the time interval of the short-time memory of human perception.

4. Experiment

We designed the continuous density HMM with a reduced number of acoustical observations represented by TDC features. The model is evaluated on a Slovak isolated digit recognition task. The speech database consists of 12 Slovak words (digits 0–9, digits 1 and 2 are spelled-out as both "jeden" and "jedna", and "dva" and "dve" respectively) uttered by 61 speakers. The database was recorded in the Department of Telecommunications in the University of Žilina. The details about the speech database are summarised in table 1. The ASR system is designed as speaker-independent. We used a different non-overlapping sub-set of the database uttered by different group of speakers for training and testing. One HMM was created for each class.

Train and test speech database Table 1

Recognition task	Isolated Slovak digits
Number of speakers	61 (40 for training + 21 for testing)
A/D conversion	sampling frequency = 8kHz, resolution = 8bits/sample, telephone quality
Number of records	4 records per word per speaker, 12 words 48 × 61 = 2928 records in total

4.1 ASR front-end

A speech signal is analysed in frames. The analysis procedure is depicted in Figure 2 . First, the signal is pre-emphasised by the 1st–order FIR filter with $k = 0.97$. The sliding 30 ms long window is used with the 20 ms shift. That means 1 second of the signal is split into 50 frames. FFT spectrum and 23 mel-FBE are computed for each frame. The frames are grouped into blocks of 12 frames with 6 frames overlap. TDC matrix is computed from each block (Eq. 2–4). Only one quarter of the TDC matrix forms a feature vector y (i.e. acoustic observation), which consists of 50 coefficients as follows

$$y = \{y_d\}=\{C(u,v),\ u = 1, 2, ..., 10;\ v = 1, 2, ..., 5\},$$

$$d = 1, 2, ..., 50. \qquad (6)$$

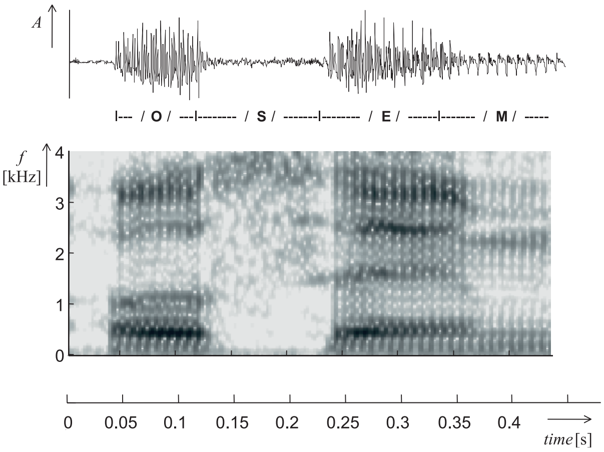

Fig. 2. Speech waveform and spectrogram of the Slovak word "osem"

Note the first row ($u = 0$) and first column ($v = 0$) of the TDC matrix are removed.

4.2 HMM Design

Let PT be the probability that a training pattern has T acoustic observations, and let T_{max} and T_{min} be maximum and minimum number of acoustic observations through all training patterns respectively. Then we propose that the number of states is given by the equation

$$N = \underset{T=T_{min}}{\overset{T_{max}}{ArgMax}} \{P_T\} \tag{7}$$

The continuous output probabilities $b_i(y)$ for each state of HMM are modelled by PDFs with a mixture of multivariate Gaussians as follows

$$p(\boldsymbol{y} \mid i) = \sum_{k=1}^{M} p(\boldsymbol{y}, k \mid i) = \sum_{k=1}^{M} p(k \mid i) p(\boldsymbol{y} \mid k, i) =$$

$$= \sum_{k=1}^{M} m_{i,k} N(y, U_{i,k}, \mu_{i,k}) \tag{8}$$

where $\boldsymbol{y}$ is the given observation, $U_{i,k}$ and $\mu_{i,k}$ are covariance matrix and mean feature vector of the k-th Gaussian component in the i-th state of HMM, and $m_{i,k}$ is the weight of k-th components. M is the number of mixture components. The term $N(y, U, \mu)$ means multivariate joint Gaussian PDF of acoustic observations $\boldsymbol{y}$, defined as

$$N(y, U, \mu) = \left[(2\pi)^{\frac{D}{2}} \sqrt{|\det U|} \right]^{-1} e^{-\frac{1}{2}(y-\mu)^T U^{-1}(y-\mu)} \tag{9}$$

where D is number of the features in one acoustic observation (D=50).

During HMM initialisation, acoustic observations of each training pattern have to be divided among N states of the model. Due to a highly reduced number of observations (see Figure 2) we proposed the following initialisation procedure: First only training utterances with the number of observations greater or equal to the number of states are selected. Their observations are allocated to the states of HMM, and centroids for each state are computed. Observations of the rest utterances are allocated to the states by a modified K–mean algorithm. Then all the observations are iteratively re-located to the states of the model by a modified K–mean algorithm. In this stage, the transition probabilities are estimated. The observations in each state are grouped to M clusters by VQ, and hyper-parameters of emission probabilities U, μ, m are estimated. A training of HMMs was performed by the well-known Baum-Welch algorithm using MLE criterion.

We tested 3 types of covariances: full, diagonal and spherical. Spherical covariance is estimated as $\boldsymbol{U} = MSE\,.\,\boldsymbol{I}$, where $\boldsymbol{I}$ is identical matrix. Mean Square Error of a cluster is given as follows

$$MSE = \underset{y}{E}\{(\boldsymbol{y} - \mu) T (\boldsymbol{y} - \mu)\} \tag{10}$$

4.3 Evaluation

ML classification using both the feed-forward and Viterbi algorithms were applied, and almost the same results were obtained for both methods (difference was only in speed of log-likelihood computations where Viterbi algorithm suits better). We evaluated PDF mixtures with 1, 2, 4, or 8 Gaussians. The results are summarised in Table 2.

The spherical and diagonal covariance gives similar results. The results, when full covariance is used, do not meet theoretical expectations (theoretically, the full covariance should be the most precise). But if we look closely at the third row of the table, we notice that the recognition rate falls down rapidly when a number of Gaussian components (and thus a number of parameters) is increasing. This effect has occurred because of an insufficient number of training data to tune all the parameters correctly (see Table 1, only 40 × 4 = 160 records per word were available for training).

Recognition rate in dependence of HMM set-up Table 2

Number of Gauss. components	1	2	4	8
	Recognition rate			
Spherical cov.	89.8 %	92.2 %	93.2 %	92.6 %
Diagonal cov.	90.6	92.6	92.3	92.3
Full covariance	87.2	66.0	20.7	9.3

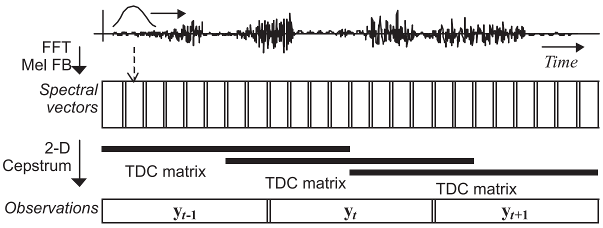

Fig. 3. Procedure of speech signal pre-processing for HMM

5. Conclusion

The ASR HMM with a reduced number of observations is proposed. In the model, one second of speech is described by only about 400 features what is ten times less than in conventional ASR systems. The performance of the system was examined on a speaker-independent isolated digit recognition task. The satisfactory results as seen in Table 2 are for the HMM with both the spherical and diagonal covariance. The best recognition rate is 93.2%. We suppose that the recognition rate will further increase if a bigger amount of training data is available. We believe that if we re-train the model on a much larger database (1000–2000 speakers is common for development of speaker-independent ASR) the model will be competitive with the ASR systems that use conventional methods. A great advantage of the proposed model is ease of computation (particularly for spherical covariance) and very fast signal processing and model training. It is suitable for application with a limited computation performance and power (e.g. mobile devices).

References

[1] O'SHAUGNESSY, D.: *Interacting with computer by voice: Automatic Speech Recognition and Synthesis,* Proceedings of the IEEE, Vol. 91, No. 9 (2003) 1272–1305.

[2] HERMANSKY, H.: *Should recognizers have ears?*, Speech Communications 25 (1998) 3–27.

[3] FURUI, S.: *On the role of spectral transition for speech perception,* J.Acoust.Soc.Am. 80(4), (1986) 1016–1025.

[4] HERMANSKY, H., MORGAN, N.: *RASTA processing of speech,* IEEE Trans. Speech Audio Process. 2(4), (1994) 578–589.

[5] JARINA, R.: *Kepstrálno-spektrálny model pre rozpoznávanie rečových signálov,* dissertation, University of Žilina (1999).

[6] KANEDERA, N, ARAI, T., HERMANSKY, H., MISHA, P.: *On the importance of various modulation frequencies for speech recognition,* Proc. Eurospeech'97, Rhodos, Greece (1997) 1079–1082.

[7] JARINA, R.: *Study of discriminative properties of two-dimensional cepstrum analysis for speech recognition,* Proc. RADIOELEKTRONIKA' 99, Brno, Czech, (1999) 168–171.

[8] ARIKI, Y., MIZUTA, S., NAGATA, M., SAKAI, T.: *Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum,* Proc. IEE, Vol. 136, Pt.I, No.2, (1989) 133–140.

[9] MILNER, B.P., VASEGHI, S.V.: *Speech modelling using cepstral-time feature matrices and Hidden Markov Models,* Proc. of the IEEE conf. ICASSP '94, Vol.I, Adelaide, Australia (1994) 601–604.

[10] KANEDERA, N., HERMANSKY, H., ARAI, T.: *Desired characteristics of modulation spectrum for robust automatic speech recognition,* Proc. of the IEEE conf. ICASSP'98, Seatle, USA (1998).

[11] JANČOVIČ, P, MACHO, D., NADEU, C., ROZINAJ, G.: *Feature selection in cepstral-time matrices for clean and noisy speech recognition,* Proc. TEMPUS-TELECOMNET workshop ITTW'98, Barcelona, Spain, July (1998) 28–36.