PSEUDO 2-DIMENSIONAL HIDDEN MARKOV MODELS IN SPEECH RECOGNITION

Steffen Werner, Gerhard Rigoll

Department of Computer Science, Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
{werner, rigoll}@fb9-ti.uni-duisburg.de

ABSTRACT

In this paper, the usage of pseudo 2-dimensional Hidden Markov Models for speech recognition is discussed. This image processing method should better model the time-frequency structure in speech signals. The method calculates the emission probability of a standard HMM by embedded HMMs for each state. If a temporal sequence of spectral vectors is imagined as a spectrogram, this leads to a 2-dimensional warping of the spectrogram. This additional warping of the frequency axis could be useful for speaker-independent recognition and can be considered to be similar to a vocal tract normalization. The effects of this paradigm are investigated in this paper using the TI-Digits database.

1. INTRODUCTION

Hidden Markov Models (HMMs) represent a well-known statistical pattern recognition technique and can be considered as the state-of-the-art in speech recognition. This is due to excellent time warping capabilities, the effective self-organizing learning capabilities and their ability to perform recognition and segmentation in one single step.

But the use of HMMs is not only restricted to time-sequential data, such as speech, handwriting or gestures. Also every static data, such as images, can be modeled and processed. In this way, images can be warped in vertical and horizontal direction, leading to elastic matching capabilities for images using HMMs. This possibility of warping in two dimensions initiates the deliberation how it is possible to use that in speech recognition, where the time sequence and the implied generation of feature sequences is obvious. This question has also been investigated recently in [2, 3].

This paper investigates the usage of image processing techniques, in order to model speech much more flexible than with standard HMMs. Therefore the feature extraction was adapted in a way that image modeling techniques could be used.

After the description of the deliberations which resulted in this work (Section 2), a brief introduction into Pseudo 2-dimensional HMMs is given in Section 3 and the adapted feature extraction and the used database are introduced in Section 4. The realized experiments and the obtained results are explained in Section 5 before a summary is presented in Section 6.

2. MOTIVATION

Standard HMMs model the temporal variability under the assumption that the speech signal is piecewise stationary. Using overlapping windows in feature extraction is due to the fact that this assumption is not realistic. The spectral variability is modeled by a Multi-Gaussian distribution which does not pay great attention to the spectral variation inside a frame.

Two-dimensional modeling techniques enable a time as well as a spectral warping possibility and consequently a modeling of the variations in both directions. Even though uncorrelated feature vector components are assumed in speech recognition, there exists always a certain degree of correlation. Therefore, the inter-speaker variation could be reduced by better modeling the correlation between the successive components of the feature vector. The larger the correlation between the components of a feature vector the better is the modeling effect in the direction of the spectral axis.

Pseudo 2-dimensional HMMs have already been used for character or document recognition [4] and face recognition [5] with good performance. With some adaptations such modeling techniques could be easily used within speech recognition, too.

3. PSEUDO 2-D HMMS

Pseudo 2-D HMMs (P2D-HMMs) are a generalization of the 1-dimensional HMM paradigm, in order to model 2-dimensional data. They are called pseudo, due to the fact that the state alignments of consecutive columns are calculated independently of each other.

As outlined in Fig. 1, P2D-HMMs, which are also known as planar HMMs, are stochastic automata with a 2-dimensional arrangement of the states. The states in horizontal direction are denoted as *super-states*, and each

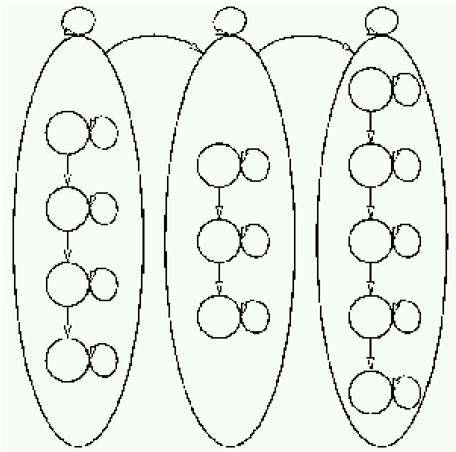

Fig. 1. Basic structure of a Pseudo 2-D HMM

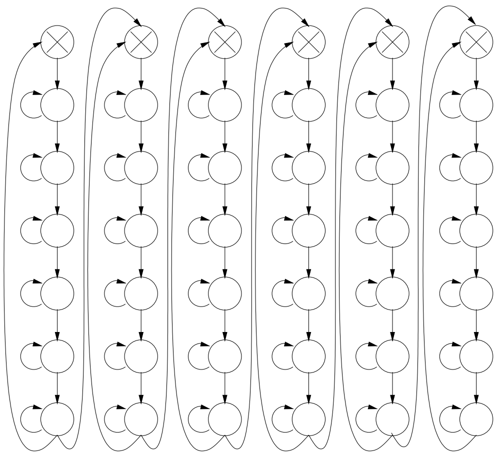

Fig. 2. Augmented 6 x 6 P2D-HMM

super-state consists of a 1-dimensional HMM in vertical direction. If one considers e.g. an image subdivided into stripes it is possible to use P2D-HMMs for modeling this 2-dimensional data in the following manner: Each stripe is aligned to one of the super-states of the P2D-HMMs, resulting in a horizontal warping of the pattern. Furthermore, within the super-state, the pattern related to the stripe is aligned to the vertical HMM states, resulting in a vertical alignment of the stripe.

In a similar way, it is possible to model data, which is considered to be consisting of horizontal stripes. How it is possible to use speech signals or parameters representing a speech signal for modeling with P2D-HMMs with respect to the advantages of this method will be discussed later (Section 4.2).

The recognition process using P2D-HMMs is similar to the recognition with "normal" HMMs, as it is accomplished by calculating the class-dependent probability that the (unclassified) data has been generated by the corresponding P2D-HMM. For this procedure, the doubly embedded Viterbi algorithm can be utilized. A detailed description of this 2-dimensional version of the Viterbi algorithm can be found in [6].

Alternatively, a P2D-HMM can be transformed into an equivalent 1-dimensional HMM as shown by Samaria in [7]. Therefore, special *start-of-line* states and *marker* features have to be inserted. Figure 2 shows an augmented 6 x 6 P2D-HMM with start-of-line states, which are indicated by a cross. When using this structure one has to take care of the fact that the value for the start-of-line feature is different from all other possible ordinary features, which forces the states to generate a high probability for the emission of start-of-line features. These equivalent HMMs can be trained by the standard Baum-Welch algorithm and the recognition step can be carried out using the standard Viterbi algorithm.

4. DATABASE AND FEATURE EXTRACTION

4.1. Database

The TI-Digits database [8] has been used throughout our experiments. The corpus contains isolated digits and sequences of up to seven digits spoken by US-American persons (men, women, boys and girls). For initialization purposes and first trainings and tests of the P2D-HMMs only the isolated digits were used.

4.2. Feature Extraction

As described in Section 2, uncorrelated features do not seem to be useful for P2D-HMMs. Therefore the experiment was performed using spectral features on the Mel-scale.

As a reference, a system based on Mel-Spectral features was used which had a vector dimension of 36 (12 Mel-Spectral, δ and $\delta\delta$). Additionally, a system was trained with 39-dimensional Mel-Cepstral feature vectors (12 MFCC, energy, δ and $\delta\delta$) in order to show the performance with more uncorrelated features.

Before using the described P2D-HMMs with standard training and recognition methods the sequence of feature vectors

$$S = \{\vec{V}_1, \vec{V}_2, \ldots, \vec{V}_i, \ldots, \vec{V}_I\} \tag{1}$$

has to be re-arranged and the *marker* features have to be added for synchronization with the start-of-line states. To make sure that the value of the marker feature is out of range, the value 20 million was chosen and added in front of every original feature vector $\vec{V}$ of a frame.

In the simplest case the re-arrangement can be done by taking the components from $\vec{V}$ as new feature vectors $\vec{F}_1, \ldots, \vec{F}_{(3 \cdot N)}$ (with N being the number of spectral feature parameters). Therefore, a warping will be done along the components of a frame. Because the δ and $\delta\delta$ features are appended to the spectral features, a simple serialization is not the best way to determine new vectors $\vec{F}$. The δ_i is much more correlated with the corresponding spectral parameter x_i, etc. Alternatively, the features could be re-arranged so that the new feature vector $\vec{F}$ includes a spectral coefficient x_i and the δ_i and $\delta\delta_i$ values. This is also the way used for our experiments here. Thus, the new feature vector sequence $\tilde{S}$ results in:

$$\tilde{S} = \{\vec{M}, \vec{F}_1^1, \vec{F}_2^1, \ldots, \vec{F}_N^1, \ldots, \vec{M}, \vec{F}_1^i, \vec{F}_2^i, \ldots, \vec{F}_N^i, \ldots, \vec{M}, \vec{F}_1^I, \vec{F}_2^I, \ldots, \vec{F}_N^I\} \quad (2)$$

The feature transformation is summarized in the following scheme.

$$\vec{V}_i = \begin{pmatrix} x_1 \\ \vdots \\ x_N \\ \delta_1 \\ \vdots \\ \delta_N \\ \delta\delta_1 \\ \vdots \\ \delta\delta_N \end{pmatrix} \Longrightarrow \begin{matrix} \vec{M} & \text{with} & \vec{M} = \begin{pmatrix} 2*10^7 \\ 2*10^7 \\ 2*10^7 \end{pmatrix} \\ \vec{F}_1^i & & \\ \vdots & & \\ \vec{F}_N^i & \text{and} & \vec{F}_j = \begin{pmatrix} x_j \\ \delta_j \\ \delta\delta_j \end{pmatrix} \end{matrix}$$

where:

$1 < i < I$ with I being the total frame number

$1 < j < N$ N being the total number of spectral parameters

$\vec{M}$ is the marker feature vector.

5. RECOGNITION EXPERIMENTS

The spoken digits were modeled by word HMMs which are strictly left to right, without skips over states. Additionally, a 3 state silence model was added which models the pause before and after an utterance. No additional pause model was used between the spoken digits of a sequence, even though it is recommended for instance in [9].

Several reference systems were trained in order to compare the results of the P2D-HMM systems. Both are explained in more detail in the following sections.

5.1. Reference Systems

As mentioned in Section 4.2 the reference systems are based on Mel-Spectral features or on Mel-Cepstral features. Both systems were trained with the whole training corpus. The recognition performance was measured on all test-files (overall performance) as well as on those files where only one digit was spoken (single digits performance). Every system was trained with up to 6 mixtures per state.

Table 1 displays the results using a 39 dimensional MFCC feature vector (12 MFCC, energy, δ and $\delta\delta$). It is obvious that this system already shows a very good performance by using only one mixture.

Table 1. *Results with Mel-Cepstral features and 10 states per model using 1D-HMMs.*

mixtures	WER (overall)	WER (single digits)
1	4.65 %	0.53 %
6	0.94 %	0.31 %

Table 2 displays the results using a 36 dimensional Mel-Spectral feature vector (12 Mel-Spectral, δ and $\delta\delta$) depending on the number of states. It is obvious that the performance decreases by decreasing the number of states. This effect can be compensated by increasing the number of mixtures.

Table 2. *Results with Mel-Spectral features using 1D-HMMs.*

states	mixtures	WER (overall)	WER (single digits)
10	1	59.63 %	14.72 %
	2	23.07 %	9.06 %
	4	12.23 %	3.29 %
	6	9.56 %	2.23 %
8	1	71.49 %	34.41 %
	6	14.23 %	2.26 %
5	1	93.35 %	47.82 %
	6	36.98 %	3.32 %

5.2. Experiments with P2D-HMMs

The P2D-HMMs are first trained and tested on files including only one spoken digit. Nevertheless, the silence model for each utterance was used, too. From a 72 dimensional feature vector (24 Mel-Spectral, δ and $\delta\delta$), 24 new feature vectors with 3 components were derived as described in Section 4.2.

Two systems were trained with 12 states on time axis (equivalent to the number of super-states). The second dimension was realized once by 10 and the other time by 5 states. The results are displayed in Table 3.

The training of both systems was completed after 2 mixtures were reached, although the result with the 12 x 10 state system is respectable. This was due to (1) the enormous increased number of parameters which have to be trained and (2) the slightly worse performance compared to all reference systems. Besides, no additional training was done

Table 3. *Results with Mel-Spectral features using P2D-HMMs, tested on files with one spoken digit.*

states	mixtures	WER
12 x 10	1	8.62 %
	2	7.19 %
12 x 5	1	42.78 %
	2	31.48 %

using the whole corpus.

The second system with 12 x 5 states shows much worse results than the other one. This might be an indication that the strong warping of frequency parameters reduces the discrimination possibilities too much.

As shown in the previous section, the decreasing of states in the time axis results in worse recognition performance which could be compensated by increasing the number of mixtures. Therefore, it should be possible to reach better results with P2D-HMMs by a smaller number of states in the time axis and a higher number of states in the frequency axis and increasing simultaneously the number of mixtures. However, it turned out that increasing the number of Gaussians did not lead to any improvements.

5.3. Analysis of the experiments

All tested systems based on P2D-HMMs have shown (partly much) worse results than the reference systems based on standard HMMs. This is probably due to the fact that discriminating characteristics are lost, such as formant frequency positions, by enabling the possibility to warp on frequency axis.

However, the best result obtained with that method (see Table 3) is comparable to the reference results obtained in Table 2 for the same number of mixtures (namely 2).

The computational effort could be reduced by using monophone models instead of word models. But as shown in [2] the results with such a system based on phoneme models are also worse with respect to the reference system there. The experiments in [2] were realized with MFCC and RASTA-PLP features.

6. CONCLUSION

In this paper an investigation on using image processing techniques for speech recognition is presented. Pseudo 2-dimensional HMMs were used to model word sequences with adapted Mel-Spectral features. Nevertheless, the speech recognition results were worse than those with conventional state-of-the-art HMMs. Only one system with 12 x 10 states could nearly reach the results achieved with standard HMMs. Regarding the enormous computational effort to train such a system compared to 1-dimensional systems, it is currently difficult to detect potential advantages of this method for speech recognition.

7. ACKNOWLEDGMENTS

The content and themes discussed in this paper largely benefits from the collaboration with the student assistant Peng Dai.

8. REFERENCES

[1] Gerhard Rigoll and Stefan Müller, "Statistical Pattern Recognition Techniques for Multimodal Human Computer Interaction and Multimedia Information Processing," in *Survey Paper, Int. Workshop "Speech and Computer"*, Moscow, Russia, Oct. 1999, pp. 60–69.

[2] K. Weber, S. Bengio, and H. Bourlard, "HMM2- A novel approach to HMM emission probability estimation," in *ICSLP*, Beijing,China, 2000.

[3] H. Bourlard, S. Bengio, and K. Weber, "New Approaches Towards Robust and Adaptive Speech Recognition," in *Advances in Neural Information Processing Systems 13*, T.K. Leen, T.G. Dietterich, and V. Tresp, Eds. 2001, MIT Press.

[4] O.E. Agazzi and S. Kuo, "Pseudo Two-Dimensional Hidden Markov Models for Document Recognition," in *AT&T Technical Journal*, Oct. 2000, vol. 72(5), pp. 60–72.

[5] Stefan Eickeler, Stefan Müller, and Gerhard Rigoll, "Improved Face Recognition Using Pseudo-2D Hidden Markov Models," in *Workshop on Advances in Facial Image Analysis and Recognition Technology in conjunction with 5th European Conference on Computer Vision*, Freiburg, Germany, June 1998.

[6] E. Levin and R. Pieraccini, "Dynamic Planar Warping For Optical Character Recognition," in *Proc. of the Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, Mar. 1992, pp. III:149–152.

[7] F.S. Samaria, *Face Recognition Using Hidden Markov Models*, Ph.D. thesis, Cambridge University, 1994.

[8] R.G. Leonard, "A database for speaker independent digit recognition," in *Proc. of the Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1984.

[9] H.G. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *Proc. ISCA ITRW for Automatic Speech Recognition (ASR)*, Paris, Sept. 2000, pp. 181–188.