COMPARISON OF WHOLE WORD AND SUBWORD MODELING TECHNIQUES FOR SPEAKER VERIFICATION WITH LIMITED TRAINING DATA

S. Euler [1] *R. Langlitz* [1] *J. Zinke* [2]

[1] Bosch Telecom
Kleyerstr. 94
D–60277 Frankfurt
Stephan.Euler@fr.bosch.de

[2]FH Giessen-Friedberg
Wilhelm-Leuschner-Str. 13
D–61169 Friedberg
Joachim.Zinke@e2.fh-friedberg.de

ABSTRACT

In this paper we use whole word and subword hidden Markov models for text dependent speaker verification. In this application usually only a small amount of training data is available for each model. In order to cope with this limitation we propose a intermediate functional representation of the training data allowing the robust initialization of the models. This new approach is tested with two data bases and is compared both with standard training techniques and the dynamic time warp method. Secondly, we give results for two types of subword units. The scores of these units are combined in two different ways to obtain word error rates.

1. INTRODUCTION

Most systems for automatic speaker verification use either the template based technique of dynamic time warp (DTW) or the hidden Markov model (HMM) framework. The HMM approach has yielded good results both for words and subwords units [1] [2]. In general this technique offers more flexibility. E.g. subword models allow extension to large vocabularies, making it almost impossible to use prerecorded speech signals to gain access. However, the amount of required training data is greater than for the DTW. In cases with very few training utterances DTW provides a better inherent time alignment resulting in lower error rates [3]. On the other hand it is much easier for HMMs to make use of more training data. The data is simply used for a better adjustment of the model parameters. In DTW increasing the number of templates results in increased storage and computation time for the verification unless techniques for merging templates are implemented.

In order to gain more insight into the effects of limited training data we compare different approaches within the continuous density HMM framework. Using a database of isolated German words we first examine for whole word models the influence of the model structure. Next we test a new approach based on a polynomial model for the words. The temporal structure of the words is then described by polynomial functions for the features. In [4] a similar approach based on an orthogonal polynomial representation for speech segments gave improvements in text–independent speaker verification.

The time continuous representation can be used directly for speaker verification. Alternatively, initial HMM parameters are derived from this representation. In this way even with limited training data a large number of states can be generated robustly. As an alternative to word models, we test phoneme-like subword units and diphone units derived from segments between the middle of consecutive phonemes. In order to compare the results two techniques for obtaining word scores from the subword models are tested.

2. FUNCTIONAL REPRESENTATION

Recent experiments with a time continuous extension of the HMM approach yielded some improvements of the recognition rate [5]. In these experiments polynomials had been used to interpolate between the HMM states. As a further step polynomials or – more general – functional approximations can be derived directly from the reference data.

For a given reference utterance Y the feature vectors $\vec{y}_t, \ldots, \vec{y}_T$ are at first assigned to time values $\tau_t, \ldots, \tau_T$. We define $\tau = -0.5$ as start of the model and $\tau = 0.5$ as end. The necessary initial values for τ can be obtained from a Viterbi segmentation of the utterance and mapping of the state sequence $Q_1, \ldots, Q_T$ into the interval $[-0.5, .5]$ in the form

$$\tau_t = \frac{Q_t - 1}{N - 1} - 0.5 \tag{1}$$

N denoting the number of states in the HMM. As an alternative, a linear increase of τ in the form

$$\tau_t = \frac{t - 1}{T - 1} - 0.5 \tag{2}$$

can be used. In this way an utterance is described by the points $(1, \tau_1), \ldots, (T, \tau_T)$ in the $t\tau$-plane. In general, a number of references $Y^1, \ldots, Y^L$ are available for each word in the vocabulary. In our approach for each feature component y the resulting set of points (y_t^l, τ_t^l) is modeled by a continuous function $\Phi(\tau)$. This

function is defined as a linear combination of M basis functions $\varphi_i(\tau)$

$$\Phi(\tau) = \sum_{i=1}^{M} c_i \cdot \varphi_i(\tau) \tag{3}$$

The quality of this approximation is given by the merit function

$$\chi^2 = \sum_{il} (y_t^l - \Phi(\tau_t^l))^2 \tag{4}$$

In our case the number of points (y_t^l, τ_t^l) is much larger than the number of coefficients. The problem of finding the optimum coefficients by minimizing χ^2 is solved by applying the singular value decomposition (SVD) method [6]. The basis functions are set to

$$\varphi_i(\tau) = \tau^{i-1} \tag{5}$$

and the approximation $\Phi(\tau)$ is a polynomial of degree $M-1$.

The set of functions $\Phi(\tau)$ for the features models the temporal structure of a word. In this way $\Phi(\tau)$ replaces the set of means $\mu_1, \ldots, \mu_N$ of an HMM with N states. Based on techniques presented in [5] the polynomials could be used directly to compute a score for a given utterance. Alternatively, the polynomials provide a robust method of initializing HMMs. The functions $\Phi(\tau)$ are per definition continuous in time allowing to generate the mean vectors for any given number of HMM states. Assuming one mixture per state, the mean for the state q_n is then given by

$$\mu_n = \Phi(\frac{n-1}{N-1} - 0.5), \quad n = 1, \ldots, N \tag{6}$$

In particular, models with a large number of states can be generated, leading to an improved modeling of temporal structures. These means can serve as initial values for further optimization. The resulting HMM in turn can be used to align the reference data. Applying the linear mapping of the state sequence new values for τ can be derived. The equations (1) and (6) allow to switch between the time continuous representation and the HMM representation. Therefore the two approaches can be applied iteratively in the training procedure allowing to take advantage of the strengths of both methods.

3. SIMULATION SYSTEM

In the simulation experiments we first used a vocabulary of 23 German words designed for applications in PABX systems (SP23). It contains the 10 digits plus the alternative version *ZWO* for the digit 2 and some command words. The speech signals were sampled at 8kHz with 8 bit logarithmic quantization and 300Hz to 3400Hz bandwidth. For each of the 23 words 10 to 15 utterances were spoken by one female and six male speakers. The first three training utterances were defined as references. In earlier experiments this data base was used for speaker recognition with DTW and VQ-HMM [7]. Comparison of different recognition systems is based on the equal error rates (EER). Therefore, the acceptance threshold is calculated individually for each speaker and word such that the number of false rejections is equal to the number of false acceptances. Best results for this data base are equal error rates of about 1% for DTW and about 4% for VQ-HMM. Additionally, a speaker independent data base with utterances of the 23 words from 200 different speakers was available for building HMMs for segmentation.

Recently, speech data has been collected during a field trial of a speaker verification system at the University of Frankfurt [8]. For more than one year access to two rooms has been controlled by a DTW based system. Wall mounted telephone sets serve for speech input. The verification system runs on a PC with a build-in DSP board and the two telephone sets are connected through a PABX system with the PC. The vocabulary consists of 8 German words (SP8). In order to gain access the user has to repeat these words, prompted in random order. After two successes, i.e. utterances with a distance score below a threshold, the user is accepted.

Collecting data from this system is an ongoing work with the aim of building up a data base with utterances from speakers over a long time period. In first experiments we used the speech data from 16 speakers for a comparison of the whole word modeling techniques. This data base has been collected over 9 month. Up to now, it includes in average 24 utterances per speaker and word.

In all simulation experiments the feature vectors consisted of LPC–based cepstral coefficients. The HMMs were trained with the HTK toolkit [9]. All models were specified with Gaussian densities and diagonal covariance matrices. The verification of an utterance Y is based on the length normalized score

$$S(Y|M) = \frac{\log(p(y|M))}{T} \tag{7}$$

of a hidden Markov model M.

4. SIMULATION RESULTS

4.1. Word models

In a first experiment we tried to find the optimum model configuration for whole word models. We used up to 7 states and 10 mixtures. The resulting EER of the 23 word data base for some configurations are shown in Fig. 1. In general, we found that for a given total number of mixtures HMMs with more states perform slightly better. But even models with one state – i.e. without any temporal structure – yield reasonable rates. The best rate of 2.3% is obtained with 7 states and 2 mixtures. This configuration was used in all following experiments. Due to the limited training data a higher number of states or mixtures increases the error rate.

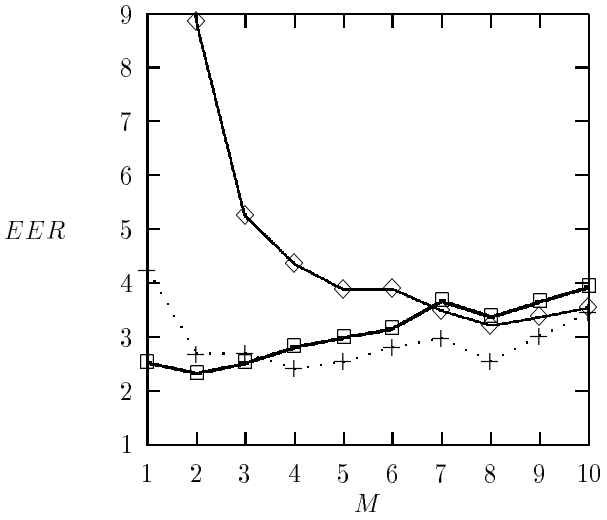

Figure 1. EER over number of mixtures, ◇ 1 state, + 4 states, □ 7 states

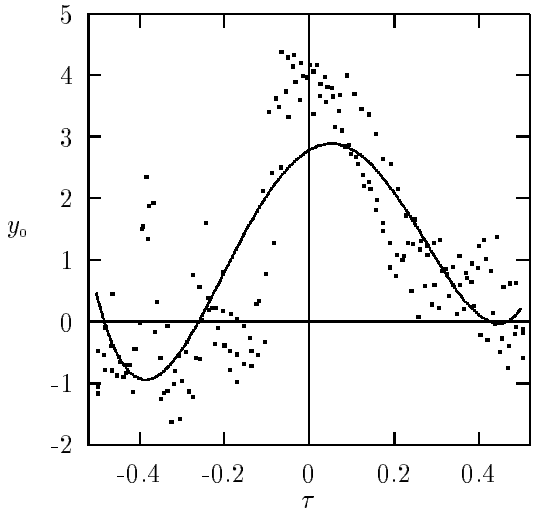

Figure 2. First cepstral coefficients of word *ZWEI* and polynomial approximation

The equal error rates vary strongly for different speakers and words. Two words (*ENDE*, *KONFERENZ*) allowed perfect verification while the highest EER was 5.0% (*RÜCKRUF*). Within the 6 speakers the EER varied between 0.2% and 6.5%.

In order to obtain robust models with more states, we tested the polynomial representation described above. At first, the feature vectors $\vec{y}_1, \ldots, \vec{y}_T$ were mapped linearly into the interval $[-0.5, 0.5]$. Next, for the resulting points (y_τ, τ), τ denoting the new time index, optimal polynomial coefficients were calculated. Each feature component is described by a polynomial $\Phi(\tau)$. Fig. 2 shows as an example the first cepstral coefficients of three utterances of the word *ZWEI* from one speaker and the resulting polynomial approximation. Using 10-th order polynomials to initialize HMMs with 18 states we achieved an ERR of 1.5%.

In Table 1 results for the two data bases and the different methods are summarized. Additionally, EER obtained with DTW are included. It should be noted that the results for DTW were obtained in simulation experiments. The installed field trial system uses an update of the templates resulting in a better performance. The new technique for obtaining initial models yields a significant improvement. In the case of the more variable data from the field trial these HMMs perform even better than the DTW system.

Comparing the different methods in detail, we found that in general the results are similar but differences between words and speakers remain. It seems therefore possible to improve the performance by combining the two approaches.

Table 1. EER for whole word models

	DTW	HMM	poly. init.
SP23	1.04%	2.32%	1.50%
SP8	12.15%	13.65%	11.73%

4.2. Subword models

In the experiments on subword models we used the 23 word data base. At first, based on a hand labeled subsection of the data base we trained speaker independent phoneme models. The 23 words in total contain 31 different phonemes. Some of the phonemes are included only once in the vocabulary. The most frequent phoneme is /n/ (as in *NEUN* /nOYn/). It occurs 14 times although the context is sometimes the same. E.g. 3 words end with the combination /@n/. Due to the small vocabulary the phoneme specific results contain a lot of context dependencies. In particular, short phonemes such as plosives tend to include a considerable part of the surrounding context into the phoneme model.

Assuming a standard transcription the speaker independent models were applied to segment the speaker dependent data and train phoneme models for each speaker. With 3 states and 3 mixtures per state we obtained a phoneme EER of 17.1%. In the next step the data was segmented again, now using the new speaker specific phoneme models. Retraining of the HMMs then yielded an improvement to 15.7%. In Table 2 rates for some typical examples are given. As expected, the vowels yield good results with an average EER of 12.7%. In general there is no clear correspondence between the frequency of a phoneme in the vocabulary and its EER. The most frequent phoneme

Table 2. EER for some phonemes

a:	E	m	N
8.2%	10.6%	8.5%	9.7%
f	v	t	k
21.0%	24.7%	25.3%	23.7%

/n/, however, also gives the highest EER of 31.0%. Furthermore, we found that plosives that occur only in one context give better results than those with multiple context.

As an alternative to the phoneme models we tested diphones that consist of the part of the speech signal between the middle of neighbored phonemes. In this way transitional parts are emphasized. Furthermore, these units already contain information about their phonetic context. For the 23 words 74 different diphones resulted. The EER using again HMMs with 3 states and 3 mixtures per state was 10.4%.

4.3. Combining subword models

In order to compare the results of the different approaches we examined two methods of combining the subword scores to word scores: addition of the phoneme scores and combining the phoneme models to a word model. In both cases always the standard transcriptions for the words was assumed. Then for the addition a word EER of 3.6% both with phonemes and with diphones resulted. It should be noted that due to the normalization of $S(Y|M)$ in (7) the contribution of the units to the word score is independent of the individual length.

Building models by concatenation of the subword models gave best results of 2.2% with phoneme models. The alternative use of diphones and additional phonemes at the word boundaries yielded 2.4%. The word HMMs in this case have a word specific number of states which is for most words much larger then in the case of whole models.

5. CONCLUSION

In this paper we first presented a new scheme for initialization of the mean vectors in HMMs from a intermediate polynomial representation. In this way even with only a small amount of training data HMMs with a large number of states can be generated robustly. Therefore this approach, allowing a better temporal modeling, leads to a significant improvement of the verification performance. In our simulation experiments with 3 reference utterances the new approach yields equal error rates close or even better than dynamic time warp. Training subword models and concatenating them to word models offers an other possibility to build large word HMMs. Our experiments showed a small improvement in the word ERR by this approach.

Combining the two techniques seems promising. Furthermore, different types of functional representations could be tested as alternatives to the polynomials. The results on the data from the field trial demonstrate that 3 references are not sufficient to cover long term variations and therefore we started to test methods for model adaptation.

ACKNOWLEDGMENT

The authors would like to thank the members of the speech recognition group at the Institut für Angewandte Physik der Universität Frankfurt and in particular H. Gruber and H. Reininger for helpful discussions and providing the speech data and DTW results from the field trial.

REFERENCES

[1] A. E. Rosenberg, C. H. Lee, and S. Gokcen. Connected word talker verification using whole word hidden markov models. In *ICASSP-91*, pages 381–384, Toronto, 1991.

[2] T. Matsui and S. Furui. Concatenated phoneme models for text-variable speaker recognition. In *ICASSP–93*, Minneapolis, 1993.

[3] K. Yu, J. Mason, and J. Oglesby. Speaker recognition models. In *EUROSPEECH 95*, pages 629–632, Madrid, 1995.

[4] C.-H. Liu and H.-C. Wang. A segmental probabilistic model of speech using an orthogonal polynomial representation: Application to text-independent speaker verification. *Speech Communication*, 18:291–304, 1996.

[5] S. Euler. A time continuous model for speech recognition. In *ICASSP–96*, pages II–889–892, Atlanta, 1996.

[6] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C.* Cambridge University Press, 1992.

[7] J. Zinke. Influence of pattern compression on speaker verification. In *EUROSPEECH 93*, pages 2267–2270, Berlin, 1993.

[8] S. Euler and H. Reininger. Zugangskontrolle durch Sprecherverifikation – erste Erfahrungen aus dem praktischen Einsatz. In *ITG–Fachbericht 139 Sprachkommunikation*, pages 85–88, Frankfurt, 1996.

[9] S. J. Young, P. C. Woodland, and W. J. Byrne. *HTK: Hidden Markov Model Toolkit V1.5.* entropic, 1993.