THE ROLE OF DYNAMIC FEATURES IN TEXT-DEPENDENT AND -INDEPENDENT SPEAKER VERIFICATION

Ying Liu, Martin Russell and Michael Carey

Department of Electronic, Electrical and Computer Engineering
The University of Birmingham, UK

y.liu@bham.ac.uk; m.j.russell@bham.ac.uk; m.carey@bham.ac.uk

ABSTRACT

A segmental hidden Markov model (SHMM) is a hidden Markov model (HMM) whose states are associated with sequences of acoustic feature vectors (or *segments*), rather than individual vectors. By treating segments as homogeneous units it is possible, for example, to develop better models of speech dynamics. This paper considers the potential benefits of a trajectory-based segmental HMM for speaker recognition. Text-dependent speaker verification (TD-SV) results obtained on YOHO and text-independent speaker verification (TI-SV) results on Switchboard are presented. The YOHO results show a 44% reduction in false acceptances using the segmental model compared with a conventional HMM, while the Switchboard results do not show any improvement relative to a conventional Gausian Mixture Model (GMM) system. Further experiments were conducted to explain these results. They indicate that the priority of a "segmental GMM" is to model stationary regions and shed light on the role of delta parameters in conventional TI-SV.

1. INTRODUCTION

Segmental hidden Markov models are intended to overcome important speech-modelling limitations of the conventional HMM approach by representing segments of features and incorporating the concept of trajectories to describe the time-varying characteristics of different speech sounds. As a first step to apply SHMMs to speaker verification, a type of SHMM with a fixed linear trajectory is used. In this type of SHMM a segment has linear trajectories each of which has a mid-point mean value and a slope to represent how the acoustic features change over time (figure 1). Each segment also has a duration probability to define the probability of segment length between 1 frame ($10ms$) and the maximum duration τ_{max}. The duration probability mass functions d_i were non-parametric [1].

This approach should have potential advantages for speaker verification. With improved modelling of speech dynamics and duration, the model should capture individual differences in non-stationary speech segments, which might otherwise be swamped by large variance due to the HMM piecewise stationarity assumption. Thus it is plausible that such a model will improve our understanding of inter-speaker differences, and hence improve speaker recognition performance, by modelling some of the underlying mechanisms that give rise to intra- and inter-speaker differences. Full details of 'Fixed Trajectory' segmental HMMs are given in [2].

This work was partly funded by EOARD (European Office of Aerospace Research and Development) award FA8655-03-1-3060 and by EPSRC grant EP/C515986/1 "A Unified Model for Speech Recognition and Synthesis".

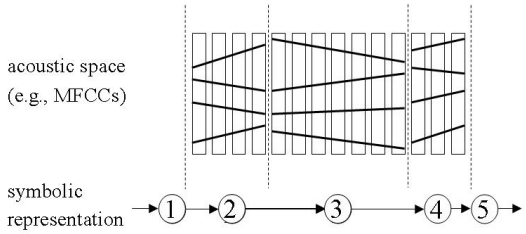


Fig. 1. A segmental HMM that uses linear trajectories and durations to represent acoustic segments.

2. THE THEORY OF SHMM

A state σ_i of a SHMM is identified with a variable duration linear trajectory representing a speech signal in a D dimensional acoustic space, which is based on Mel-Frequency Cepstral Coefficients (MFCCs). Thus a state is parameterised by the mid-point vector $\mathbf{c}_i$ and slope vector $\mathbf{m}_i$, and a $D \times D$ covariance matrix V_i. A trajectory $\mathbf{f}$ of length τ is defined by:

$$\mathbf{f}_i(t) = (t - \bar{t})\mathbf{m}_i + \mathbf{c}_i \tag{1}$$

where $\bar{t} = (\tau + 1)/2$. If $Y_1^\tau = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_\tau]$ is a sequence of acoustic feature vectors then the probability (density) of Y_1^τ given state σ_i is given by:

$$p(Y_1^\tau|\sigma_i) = b_i(Y_1^\tau) = d_i(\tau)\prod_{t=1}^{\tau}\mathcal{N}\left(\mathbf{y}_t; \mathbf{f}_i(t), V_i\right), \tag{2}$$

where $d_i(\tau)$ is the probability that state σ_i emits a segment of length τ, and $\mathcal{N}\left(\mathbf{y}_t; \mathbf{f}_i(t), V_i\right)$ is a D dimensional Gaussian probability density function (PDF) with mean $\mathbf{f}_i(t)$ and covariance matrix V_i (it is assumed that V_i is diagonal). The segmental Viterbi decoder, which is used for training and recognition, is described in [3] along with the procedures used to estimate the parameters $\mathbf{c}_i$, $\mathbf{m}_i$, and V_i.

3. TEXT-DEPENDENT SPEAKER VERIFICATION

The most straightforward application of SHMMs to speaker recognition is text-dependent speaker verification. This is because a conventional TD-SV system typically uses phone-level or word-level HMMs, which can simply be replaced by the corresponding SHMMs.

Our experiments used the YOHO [4] and TIMIT [5] speech corpora. Thirteen dimensional feature vectors were extracted comprising MFCCs 1 to 12 plus energy. No Δ or Δ^2 parameters were

used[1]. Context-sensitive triphone HMMs and SHMMs were trained using the above speech data. The conventional HMMs were constructed using the Hidden Markov Model Tool Kit (HTK) [6], and the SHMMs using the 'SEGVit' software developed at the University of Birmingham. The maximum segment duration τ_{max} of SHMMs was set to 15 ($150ms$). The detailed description of this experiment is presented in [7].

The results of the text-dependent speaker verification experiments are shown as DET curves in figure 2. The lower-bound of 0.17% in the figure for the false rejection probablity equates to a single rejection out of the 590 'authorised user' trials. Both systems achieve an optimal false rejection rate of 0.5%. The false acceptance rate for the conventional HMM system (dashed line) and SHMM system (solid line) at the optimal points are 0.52% and 0.29%, corresponding to 359 and 200 false acceptances respectively out of the 69,030 'impostor' trials. This equates to a 44% reduction in the number of false acceptances by using the SHMM system, relative to the conventional HMM-based system.

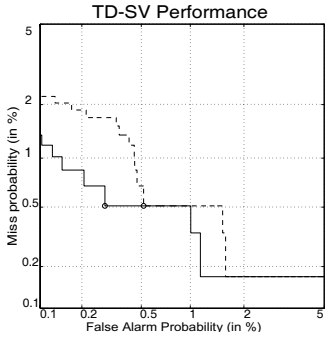


Fig. 2. TD-SV results on YOHO using HMMs (dashed line) and SHMMs (solid line).

4. TEXT-INDEPENDENT SPEAKER VERIFICATION

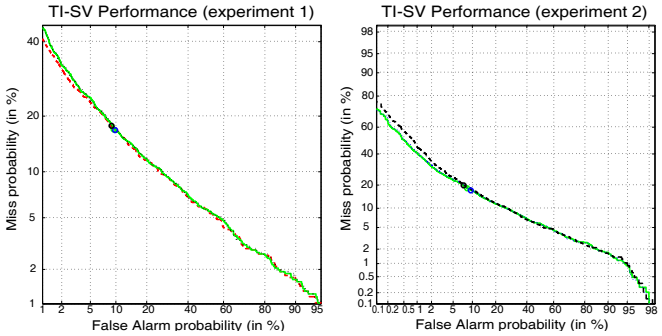
A segmental HMM version of a conventional Gaussian Mixture Model (GMM) based speaker recognition system [8] has been developed and applied to TI-SV on the Switchboard corpus. While a conventional GMM system analyses each acoustic feature vector in a speech signal separately, a segmental system attempts to model the speech signal as a sequence of variable length acoustic segments.

The experiments use the one-speaker training material from the 2002 NIST SRE to train the background model (BM), and a subset of the one-speaker data from the 2003 NIST SRE as training data for speaker-dependent model (SDM) and test data for verification. 19 dimensional feature vectors were extracted comprising MFCCs 1 to 18 plus energy. Both the GMM system and the SHMM system were constructed using the 'SEGVit' software. The GMM system was constructed with the segment durations fixed to 1 ($10ms$) and trajectory slopes set to zeroes. The number of segmental components in both GMMs and SHMMs was set to 300. Determination of the optimal segment (state) sequence uses segmental Viterbi decoding [3]. The detailed description of this experiment is presented in [7].

[1]We have not yet used Δ or Δ^2 parameters in any of our previous SHMM based experiments. This is mainly because part of the motivation for the development of SHMMs is to obtain a better model of speech dynamics and thereby obviate the need for these parameters, and also to reduce the computational load.

Two experiments were conducted: The first experiment investigated the effects on performance of setting the trajectory slope values to zero in both the BM and SDMs (SW5_1), reestimating the trajectory slope vector for the BM but not for the SDMs (so that the SDM trajectory slope vectors are equal to the corresponding BM slope vectors, SW5_2), and reestimating the slope vector for both the SDMs and the BM (SW5_3). In this experiment the maximum segment duration τ_{max} was set to 5. The second experiment constructed three systems with maximum duration τ_{max} equal to 1(SW1), 5(SW5) and 10(SW10). In this experiment all of the BM trajectory parameters were reestimated and used to seed the corresponding SDM parameters, and all of the SDM parameters were then reestimated (except in the case of SW1, where the slope vectors are all zero - this is the baseline system).

The results for both experiments are shown as DET curves in figure 3. The figure shows that the equal error rate for all systems in both experiments is approximately 14%. In the first experiment (figure 3a) the best performance is obtained using speaker-dependent trajectory slopes (scheme 3 - dashed line), but the difference between this and the other results (trajectory slopes set to zero (scheme 1 - dotted line), trajectory slopes reestimated for the BM but not reestimated for the SDMs (scheme 2 - grey solid line)) is very small and unlikely to be significant. Similarly, in the second experiment (figure 3b) the systems with $\tau_{max} = 5$ (scheme 2 - dotted line) and $\tau_{max} = 10$ (scheme 3 - grey solid line) work very slightly better than the system with $\tau_{max} = 1$ (scheme 1 - dashed line), but still the differences are too small to be significant.



a. Experiment 1 b. Experiment 2

Fig. 3. TI-SV Results on Switchboard using GMMs and SHMMs.

All of these results are clearly much worse than the best performance obtained by Lincoln Laboratory at Massachusetts Institute of Technology on the full 2003 test set using a conventional GMM system, which is a little over 5% equal error rate. This was obtained using a 2048 component GMM system, T-norm, a support vector machine and a biologically inspired acoustic parameterisation.The poor performance of our system is due to smaller number of components, the absence of normalisation and different front-end analysis.

5. ANALYSIS ON TI-SV SYSTEM

We conducted a set of experiments to investigate whether the trained SHMMs successfully contain speech dynamics and if so, whether this information can contribute to speaker verification performance. Several different parameter sets were used to train the BM and the

trained BMs were analysed.

5.1. Segment slopes of BM trained on Switchboard

Given the results in figure 3, an obvious question is whether the SHMM is actually capturing dynamic information. An initial analysis of the values of the slopes in this system suggests that it is not; the majority of slopes are close to zero. To understand why this is the case, we focused on the dynamic behavior of individual, or reduced sets of MFCCs.

Ten 300-segment BMs were trained on different sets of MFCCs. In each set a different number (from 1 to 10) of MFCCs including MFCC_0 were used. For example, in the first set only MFCC_0 was used, and in the tenth set MFCC_1 to MFCC_9 and MFCC_0 were used. Figure 4 shows the MFCC_0 slopes of different BMs. For clarity only 5 of the 10 sets are shown on the graph. As the number of MFCCs increases, the percentage of non-zero slopes decreases. This suggests that the lack of non-zero slopes is due to the maximum likelihood training algorithm giving priority to modelling stationary regions, together with the combinations of modelling these regions for all of the MFCC parameters. If this is the case, one would expect to see more non-zero slope values if the number of segments is increased.

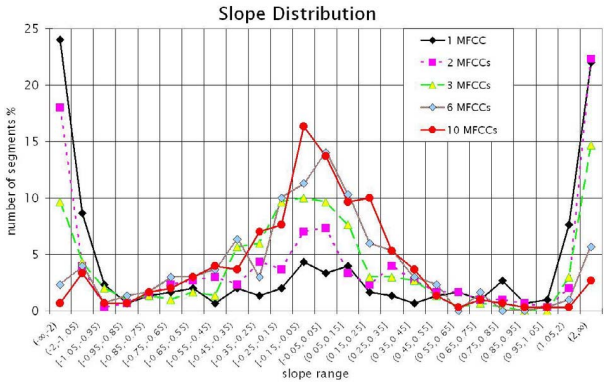


Fig. 4. Slopes of MFCC_0 in BM as number of MFCCs increases.

The second experiment was to observe the effect of different number of segments on trained BM slopes. We fixed the number of MFCCs at 6 (MFCC 1 to 5 plus MFCC_0), and increased the number of segments in SHMMs from 300 to 2100, with the intervals equal to 300. As predicted the analysis on trained BMs shows that as the number of segments increases, more percentages of segment trajectories tend to have bigger slopes. Figure 5 shows the changes of MFCC_6 slopes as the number of SHMM segments increases. The MFCC_6 slopes of the BMs with 300 and 2100 segments are drawn as solid curves with circle and square marks, separately. MFCC_6 slopes of other BMs are shown as dashed curves. In figure 5 the slope distribution of MFCC_6 in one slope range $[-0.05, 0.05]$ is exceptional, in which the percentage of segments first decreases as the number of segments increases but then it increases again as the number of segments gets bigger than 1500.

5.2. Comparison of dynamic information contained in GMM and SHMM system

As previously described, there are no Δ or Δ^2 parameters in our SHMM system because we hope to represent acoustic dynamics by

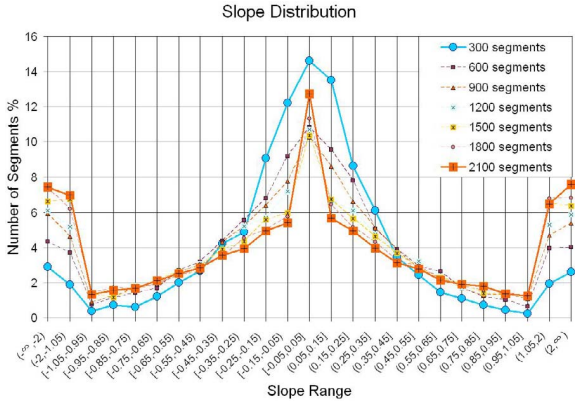


Fig. 5. Slopes of MFCC_6 in BM as number of segments increases.

using segment trajectories. We constructed a SHMM system with a duration length τ set to 2. The segment slopes were then analysed and compared with the 'delta' parameters in a traditional GMM-based system. A traditional GMM background model trained on Switchboard[2] was compared with our SHMM background model. Figure 6 compares the distribution of slope values in our SHMM and delta values in the conventional GMM. Surprisingly, the 'delta' parameters in GMMs are even smaller, 48.3% of the GMM 'delta' parameters are distributed in the range $[-0.05, 0.05]$ compared with 28.9% of the SHMM trajectories.

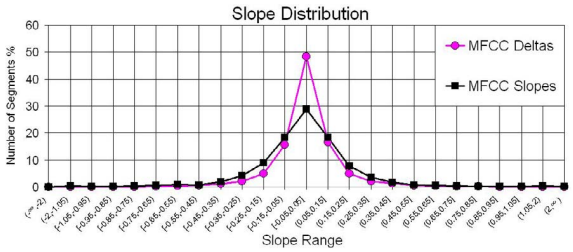


Fig. 6. Statistics of GMM deltas and SHMM segment slopes.

The result indicates that the absence of a model of dynamics in TI-SV is not just a feature of our SHMM system. It is also a feature of a conventional TI-SV GMM. This suggests that the role of delta features in such a system is not to model varying dynamics but to focus the modelling onto the stationary regions of a speech signal.

6. ANALYSIS OF TD-SV SCORES ON YOHO

We then returned to the YOHO results to look for evidence that the improvement on TD-SV scores is due to better representation of dynamics by SHMMs. Statistics of BM slopes (figure 7) show that the YOHO SHMM slopes are more diverse than the Switchboard system. Less than 9% of the segment slopes are distributed in the range $[-0.05, 0.05]$.

Do these dynamic regions contribute to speaker-detection accuracy? Experiments were conducted to find out if there is any correlation between the contribution to the YOHO speaker verification

[2]The authors wish to thank Eric Hansen and Tim Anderson from the Air Force Research Laboratory, for providing a conventional background GMM [9].

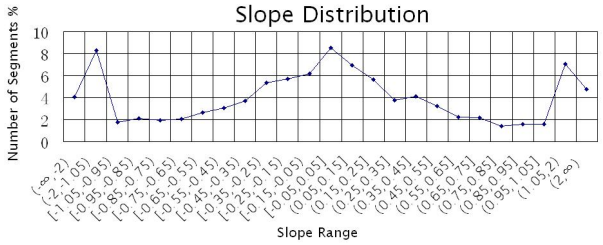


Fig. 7. Statistics of SHMM segment slopes in YOHO.

score due to a segment and the SHMM trajectory slopes for that segment. By measuring the likelihood ratio $p(Y|S)/p(Y)$ for individual segments Y of a speech signal, we could find out the relative contributions of static and dynamic segments to the speaker-detection decision.

The segment-level scores were extracted and compared with the BM segmental trajectory slopes. The average scores for each segment over all test samples of this segment were extracted. Then the average scores of all segments from each slope range were calculated. In total 127 context-sensitive triphone SHMM states were used in the YOHO TD-SV system. All SDM scores have been normalised by the BM scores in the logarithmic domain and normalised by segment durations. The sum of all 19 MFCC slopes (absolute values) in each segment was calculated to show the "non-stationarity" of each segment. The bigger the true speaker scores, or, the smaller the impostor scores, the better the contribution of the segment to TD-SV. A baseline system was also built for a reference in which all the triphone models have zero slopes.

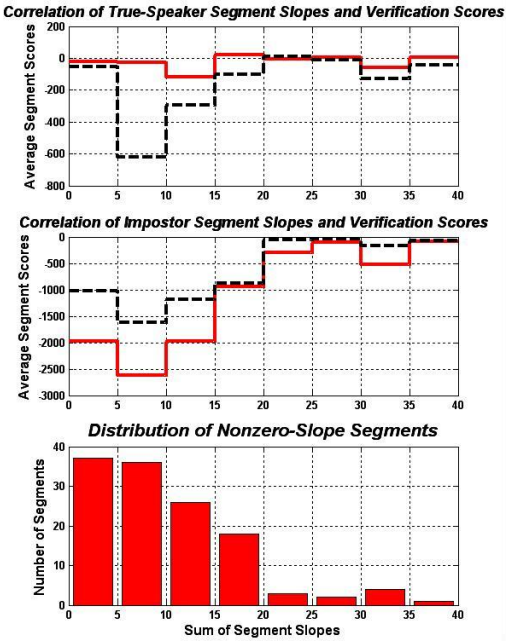


Fig. 8. Comparison between the TD-SV scores of the nonzero and zero -slope segments. (*solid line - trajectory slope SHMMs; dashed line - zero slope SHMMs).*

Figure 8 shows the comparison of both systems. Although in the baseline system all the segments have zero slopes, to make a clear comparison with the non-zero slopes, the same slope distribution of nonzero-slope BM segments was used to locate the baseline scores. The trajectory slope SHMMs were represented as the solid line. The zero-slope SHMMs were represented as the dashed line. The number of segments in each slope range was also displayed. The analysis shows that the nonzero-slope segments have bigger true speaker scores and smaller impostor scores. The increases of true speaker scores are most significant in the slope range from 5 to 20 and the decreases of impostor scores are most significant in the slope range from 0 to 15. Both areas contain most of the segments. Thus, the SHMMs in a TD-SV system do contain speech dynamic information and from our analysis these dynamic regions do contribute to speaker verification accuracy.

7. CONCLUSIONS

This paper uses trajectory-based segmental models to explore the role of dynamic information in text-independent and text-dependent speaker verification. The results indicate that in a TD-SV system, based on phone-level models, dynamic structure is exploited to improve performance. This is because the requirement to model explicit phone-level units encourages the models to take account of non-stationary regions. However, for a TI-SV system based on "segmental" GMMs, the priority of the maximum likelihood training algorithm appears to be to model stationary regions, and the trajectory slopes are simply used to focus on these regions. Our analysis of a conventional GMM system confirms this view - the role of the delta parameters in a conventional GMM-based TI-SV system appears to be to ensure that classification focuses on stationary regions rather than to model arbitrary dynamic regions.

8. REFERENCES

[1] Ferguson, J. D., "Hidden Markov Analysis", in *Hidden Markov Models for Speech*, Institute for Defense Analyses, Princeton, NJ, 1980.

[2] Holmes, W. J. and Russell, M. J., "Probablistic-trajectory segmental HMMs",Comp. Speech & Lang.,Vol.13(1),3-37,1999.

[3] Russell, M. J. and Jackson, P. J. B., "A multiple-level linear/linear segmental HMM with an 'articulatory' intermediate layer", to appear in Comp. Speech & Lang.

[4] Higgins, A., "YOHO Speaker Verification", Presented at the Speech Research Symposium, Baltimore, MD, 1990.

[5] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, 1993.

[6] Young, S. J. and Odell, J. and Ollason, D. and Valtchev, V. and Woodland, P., "The HTK Book", Version 2.1, Entropic Camb. Res. Lab., Cambridge, UK, 1997

[7] Liu, Y. and Russell, M. J., "Speaker recognition using a Trajectory-Based Segmental HMM", proceedings of Odyssey 04, the speaker and language recognition workshop, 2004.

[8] Reynolds, D. A., "A Gaussian mixture modeling approach to text independent speaker identification", PhD Thesis, Georgia Institute of Technology, 1992.

[9] Hansen, E. G., Slyh, R. E. and Anderson, T. R., "Speaker Recognition using Phoneme-Specific GMMs", proceedings of Odyssey 04, the speaker and language recognition workshop, 179-184, 2004