The Role of ‘Delta’ Features in Speaker Verification

Ying Liu, Martin J. Russell, Michael J. Carey

Electronic, Electrical and Computer Engineering, University of Birmingham, UK
y.liu@bham.ac.uk, m.j.russell@bham.ac.uk, m.carey@bham.ac.uk

Abstract

Our previous experiments in Text-Dependent and -Independent Speaker Verification (TD-SV and TI-SV) using trajectory-based models, showed that non-stationary segments benefit TD-SV but not TI-SV, because in TI-SV maximum likelihood (ML) training results mainly in stationary segments. This result questions the role of non-stationary, ‘delta’ parameters in conventional GMM-based TI-SV. In this paper we develop and study a number of GMM-based TI-SV systems for Switchboard which use combinations of static and dynamic parameters. We show that in our segmental GMM and the AFRL GMM system, the trajectory slopes and deltas focus the verification process onto the stationary regions. In our GMM systems, however, the deltas are modelling some speech dynamics. The different functions of deltas may be due to different system settings and front-end processing (e.g. RASTA, speech noise detector). This indicates that the role of delta parameters in GMM-based speaker verification systems is more complex than simply “modelling dynamics”. Our results also show that the superior performance obtained with front-end parameterizations which combine static and delta parameters only emerges after RASTA filtering; without RASTA filtering a ‘delta-only’ front-end performs best.

Index Terms: speaker verification, speech dynamics, delta parameters, segmental HMMs, GMMs.

1. Previous research using SHMMs

The goal of the work described in this paper is not, explicitly, to improve speaker verification performance. Instead the goal is to understand one aspect of how current GMM-based speaker verification systems work; namely the role of dynamic ‘delta’ parameters. Of course, this in turn may lead to performance improvement. In previous research [1, 2], trajectory-based SHMMs were used to explore the utility of modelling speech dynamics in Text-Independent and Text-Dependent Speaker Verification. Our SHMMs use linear trajectories to represent how the acoustic features change over time.

Results on YOHO showed that this type of SHMM outperforms HMMs in TD-SV, but in TI-SV on Switchboard the segmental GMMs with non-zero slopes gained no significant improvement over segmental GMMs with zero slopes. Further analysis showed that in a TD-SV system the use of explicit phone-level models ensures that dynamic structure is modelled and exploited to improve performance. However, for a TI-SV system based on segmental GMMs there is no such constraint. An analysis of the TI-SV system showed that the number of segments with non-zero slopes decreases as the acoustic feature vector dimension increases, or as the number of segments decreases. In this case 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. The complete analysis is presented in [2].

Figure 1 compares the distribution of the trajectory slope values in the cases where the ‘background’ model is a ‘segmental GMM’ (used for TI-SV on Switchboard), phone-level SHMMs (used for TD-SV on YOHO), and the distribution of the delta values in a conventional ‘background’ GMM (again used for TI-SV on Switchboard) from AFRL (see 6). These trajectory slopes and deltas are from all components of all the MFCC vectors. The figure illustrates that the slope values for the two TI-SV models are concentrated around zero. In the TD-SV models, a larger proportion of the slope values are significantly non-zero, which indicates a greater emphasis on modelling speech dynamics. Surprisingly, perhaps, the delta parameters in the AFRL GMM have the smallest values, with 57.1% of the GMM delta parameters distributed in the range of $[-0.05, 0.05]$, compared with 28.9% of the Switchboard ‘segmental GMM’ trajectories. The YOHO phone-SHMM slopes are most diverse of the three. Less than 9% of the segment slopes are distributed in the range around zero. It is shown in [2] that in the TD-SV case the non-zero trajectory slopes contribute to verification accuracy.

2. TI-SV using conventional GMMs

Previous research [4, 5] has concluded that using MFCC delta parameters alone in TI-SV leads to much poorer performance compared with either using static parameters alone or static plus delta parameters. To investigate how traditional GMM systems handle dynamics and how ML training deals with delta parameters in a GMM system, we build three traditional GMM systems using different parameter sets: sys_19 (19 static parameters (MFC_0 to MFC_18)), sys_19d (19 deltas (Δ MFC_0 to Δ MFC_18)), and sys_38sd (19 statics plus deltas).
2.1. Experiment methods

After the front-end processing the MFCCs were extracted. We applied Cepstral Mean Subtraction (CMS) over each speech file to remove possible convolutional noise due to channel effects. Then an energy-based speech-noise detector was used to judge which parts of the speech are noise and which are speech. The noise parts were discarded to remove irrelevant information.

Three GMM systems were built each of which has a different parameter set. Their background models, or World Models (WMs) were trained using the NIST 2002 SID one-speaker training material. Each WM was initialized with a single mixture component with a global mean and variance. The model components were then repeatedly split and reestimated until each of the GMMs contained 512 Gaussian mixture components, which was the maximum number of components that could be well trained using the 2002 material.

The Speaker Models were obtained from the WMs by MAP adaptation [6] using half of the one-speaker data from the NIST 2003 SRE training set. Accordingly half of the 2003 NIST SRE test data were used. The chosen subsets of training and test data were exactly the same as those used in our segmental GMM system. During MAP adaptation, for each component only the model means were re-estimated, according to Equation (1):

\[
\bar{\mu} = r \cdot \mu_b + (1 - r) \cdot \mu_W,
\]

where \(\bar{\mu}\) is the new (MAP adapted) value of the mean, \(\mu_b\) is the mean of the speaker adaptation data, \(\mu_W\) is the mean from the World Model, and \(r = n/(n + R)\), where \(n\) is the number of occurrences of the current component as the best scoring mixture component. Following Carey’s work [7], \(R\) was set to 16. The Hidden Markov Model Tool Kit (HTK) [8] was used to train the GMMs and for the speaker verification trials.

We used test normalization (T-norm) [9] for score distribution scaling. During test a set of example impostor models was used to calculate log-likelihood scores for each utterance. From these scores a mean and variance were estimated. All verification scores for the utterance were normalized by subtracting this ‘impostor mean’ and dividing by the ‘impostor variance’.

2.2. Results

The Speaker Verification performances of the three GMM systems are shown as DET curves in figure 2. Surprisingly the best performance, with an EER around 10%, was achieved by the ‘delta-only’ system (sys_19d), which is an EER between 9% and 10%. The ‘static-only’ system, (sys_19s), achieves an EER of approximately 15%, and the system which employs both static and delta parameters has a performance better than the ‘static-only’ system but worse than the ‘delta-only’ system, with an EER around 12%. This pattern is clearly different from other published results, which typically show that the ‘static-plus-delta’ system works better than the ‘static-only’ system, and that both systems work better than the ‘delta-only’ system. We tried many different system settings, but the ranking of the results remained the same. The ‘delta-only’ system always gave the best performance.

A remaining difference between our system and others which perform well for TI-SV on Switchboard and achieve best results using ‘static-plus-delta’ parameters, is that our system does not include any provision for noise robustness other than CMS. This led us to speculate that the superior performance of our ‘delta-only’ system was due to the robustness of the ‘delta’ parameters to noise.

3. RASTA filtering

One of the most popular approaches to noise robustness is RASTA filtering [10]. RASTA filtering is motivated by the observation that human hearing is relatively insensitive to a slow change in the frequency characteristics of the communication environment and thus steady background noise does not severely impair human perception of speech. RASTA filtering uses a spectral estimate in which the time varying signal in each frequency channel is band-pass filtered. Hence RASTA filtering suppresses components of the time-varying signal in each channel which change too quickly or too slowly.

3.1. RASTA filtering and speech dynamics

A typical implementation of RASTA filtering begins with the transformation of the speech signal into a regular sequence of short-term critical-band log spectra. Next, for each spectral channel the temporal derivative is calculated using a regression line through five consecutive time values (50ms). The sequence of spectra is then recomputed by integrating these derivatives through time in each spectral channel. This whole process is equivalent to first-order IIR filtering of each frequency channel time series. It is common practice to apply RASTA filtering to the sequence of MFCC vectors, rather than log spectral vectors, and this was done in the current experiments. The new static features, which were estimated using these dynamic features, will be less sensitive to both very slow variations and very fast frame-to-frame variations in the short-term cepstrum.

3.2. Experiment results

The results after RASTA filtering are shown in figure 3. In these experiments the systems used CMS and variance normalization over each 3 second speech segment, an energy-based speech-noise detector and T-norm. The energy component, MFC_0, was discarded as a form of energy normalization. The results show a substantial improvement in the performance of the ‘static-only’ system, with a new EER of 10%. The performance of the ‘delta-only’ system is slightly worse, having an EER just above 10%. The ‘static-plus-delta’ system now gives the best performance, with an EER around 8%. The shift of system performances supports the hypothesis that the superior perfor-
mance of the ‘delta-only’ system prior to RASTA filtering is due to the noise-robustness of the deltas.

3.3. Further analysis of the delta parameters

Figure 4(a) shows the cumulative distributions of the absolute delta values for the 3 systems from figure 1 and for the ‘delta-only’ and the ‘static-plus-delta’ GMM systems. Figure 4(b) zooms in to show the probability of the delta value up to 0.1.

The figure shows that the delta parameters are being used in quite different ways in these systems. For the AFRL GMM and segmental GMM systems, a large proportion of the deltas and slope values are close to zero, suggesting that the role of these parameters is to focus the verification process onto the stationary regions. In contrast, the ‘static-plus-delta’ system has a large proportion of non-zero delta parameters, suggesting that they are actually modelling dynamics. In figure 4(b) we can see that although the ‘static-plus-delta’ system contains more non-zero deltas than the AFRL GMM and segmental GMM systems, around 15% of its deltas are equal to or smaller than 0.01, and nearly 10% of its deltas are zeroes. The local optimality of the EM algorithm is a possible reason that these systems end up with different delta values after training. Alternatively, this may be due to other differences between these systems, such as the use of speech/noise detector or CMS. Whatever the explanation, the results show that the role of delta parameters in GMM-based speaker verification systems is more complex than simply “modelling dynamics”.

4. Fusion of the ‘delta-only’ and ‘static-plus-delta’ systems

The fact that the delta values are different in the ‘delta-only’ and ‘static-plus-delta’ systems, plus the fact that the ‘delta-only’ system performs surprisingly well, suggest that it may be advantageous to base the verification decision on a combination of the scores from the two systems. Consequently we investigated a ‘fused’ score function of the following form:

\[ S_f = \lambda \cdot S_{s+d} + (1 - \lambda) \cdot S_d, \]  

where \( S_{s+d} \) is the ‘static-plus-delta’ score, \( S_d \) is the ‘delta-only’ score, and \( S_f \) is the fused score. The fusing parameter \( \lambda \) is an empirically determined value between \([0, 1]\).

4.1. Fusion results

The result of fusing the ‘static-plus-delta’ and ‘delta-only’ scores, before T-norm is applied to either score, is shown in figure 5. The fused system works better than either the ‘statics-
fusing the two systems after T-norm has been applied to both of them. In this case fusion appears to offer no advantage. This suggests that, at least after T-norm, the scores of the ‘delta-only’ and ‘static-plus-delta’ systems are correlated.

4.2. Correlation of delta-only and static-plus-delta scores

Figures 7(a) and (b) show scatter plots of the ‘delta-only’ score against the corresponding ‘static-plus-delta’ score, before and after application of T-norm, respectively. In both cases the scores from both systems are clearly correlated, as one would predict from the small effects gained by fusing the two.

5. Conclusions

The experiments in this paper were motivated by our earlier experiences of applying trajectory models to TI- and TD-SV. Briefly, the segmental work showed that in TD-SV, trajectories were being used to model non-stationary regions of speech patterns and were contributing to improved performance. In contrast, in TI-SV on Switchboard there was no improvement and the trajectory slopes were all close to zero after ML training.

This led us to question the role of delta parameters in conventional GMM-based TI-SV systems for Switchboard. We show that in our segmental GMM and the AFRL GMM system, the trajectory slopes and deltas focus the verification process onto the stationary regions. In our GMM systems, however, the deltas are modelling some speech dynamics. The different functions of deltas may be due to the local optimality of the EM algorithm, the different system settings (e.g. the number of states) and the front-end processing (e.g. RASTA, speech noise detector, CMS). This indicates that the role of delta parameters in GMM-based TI-SV systems is more complex than simply “modelling dynamics”.

Our results also show, surprisingly, that prior to RASTA filtering the system based on deltas alone outperforms corresponding systems based on ‘static-only’ or ‘static-plus-delta’ parameters. However, after RASTA filtering the ordering is reversed, with the ‘static-plus-delta’ system performing best and the ‘delta-only’ system performing worst. This suggests that the good performance of the ‘delta-only’ system may be due to their tolerance to noise (on which RASTA filtering relies) rather than their ability to capture speech dynamics.

Finally, we have shown that the scores produced by the ‘delta-only’ and ‘static-plus-delta’ systems are correlated, and that there is little to be gained by fusing the two systems.

6. Acknowledgments

This work was funded by EPSRC grant EP/C515986/1 “A Unified Model for Speech Recognition and Synthesis”.

The authors wish to thank Eric Hansen and Tim Anderson from the Air Force Research Laboratory (AFRL), for providing the conventional background GMM used to produce figure 1.

7. References