Combining Spectral and Prosodic Features in HMM-based Single Utterance Speaker Verification

Osman Büyük[1] and Levent M. Arslan[2,3]
[1]*Electronics and Communications Eng. Dept., Kocaeli University, Kocaeli, Turkey*
[2]*Electrical and Electronics Eng. Dept., Bogazici University, Bebek, Istanbul, Turkey*
[3]*Sestek Inc., ITU Ayazaga Kampusu, ARI-1 Teknopark Binasi, Maslak, Istanbul, Turkey*

Keywords: Speaker Verification, Text-Dependent, Single Utterance, Sentence Hmm, Prosodic Features.

Abstract: In this paper, we combine spectral and prosodic features together in order to improve the verification performance on a text-dependent single utterance speaker verification task. The baseline spectral system makes use of a whole-phrase sentence HMM topology for the fixed utterance. We extract prosodic features using time alignment information obtained from the HMM states. In our experiments we observe that, although the prosodic features individually do not yield high performance, they provide complementary information to the spectral features. We achieve approximately 10% relative reduction in EER when the information sources are combined with a multi-layer neural network.

1 INTRODUCTION

Speaker recognition is the task of recognizing a person from his/her voice. Most speaker recognition systems rely on the low-level information via short-term spectral features. However, especially in text-independent applications, when relatively large amount of speech is available from a speaker, high-level information sources (e.g. prosodic, lexical, phone and conversational features) have provided complementary information to the spectral features (Weber et al., 2002, Reynolds et al., 2003, Klusacek et al., 2003, Dehak et al., 2007, Shriberg, et al., 2005, Ferrer et al., 2010). Moreover, they are known to be less susceptible to channel variations and background noise.

Over the recent years, much of the effort in speaker recognition research has been concentrated on text-independent applications. This can be mainly attributed to almost annual NIST evaluations (NIST, 2012). On the other hand, text-dependent applications have gained more attention in private sector for fraud prevention because of ease of use and higher accuracy for relatively short enrollment and authentication utterances. Those systems also offer significant cost reduction in call centers since they can reduce or eliminate the need for identity check questions.

One of the main objectives of a practical call center verification application is to reduce the processing time required to authenticate a user. In this kind of application, the data amount may not be adequate to reliably estimate high-level feature parameters. However, if the enrollment and authentication sessions involve the repetition of a single utterance, these features may provide additional gains. In this paper, we combine spectral and prosodic features together in a text-dependent single utterance (TDSU) speaker verification task. For this purpose, we collected a multi-channel speaker recognition database which consists of multiple recordings of a single Turkish sentence. In the future, we plan to distribute the database for academic research purposes.

Previously in (Yegnanarayana et al., 2005), spectral, source, and suprasegmental (pitch and duration) features are combined in a TDSU task. In the study, the baseline spectral system makes use of the dynamic time warping (DTW) technique. Suprasegmental features are extracted using the warping path information in the DTW algorithm. In (Charlet et al., 2000), discrete state duration modeling is used in a HMM-based framework. However, other prosodic features such as pitch and energy are not tested in this study. Different from the previous studies, we use whole-phrase sentence HMM structure as the baseline spectral system and

Büyük O. and M. Arslan L..
Combining Spectral and Prosodic Features in HMM-based Single Utterance Speaker Verification.
DOI: 10.5220/0005217500860091
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (BIOSIGNALS-2015), pages 86-91
ISBN: 978-989-758-069-7

extract duration, pitch and energy statistics using the time alignment of the HMM states. We fuse the scores of the spectral and prosodic systems using a three-layer perceptron network. Additionally, test normalization (T-norm) and handset-dependent test normalization (HT-norm) (Auckenthaler et al., 2000) are applied on the final likelihoods in order to reduce the effects of channel mismatch.

The remainder of this paper is organized as follows. In Section 2, we provide the details of our speaker recognition database. In Section 3, our methodology is presented. Section 4 is devoted to experimental setup and verification results. We conclude the paper with a summary of results and observations.

2 DATABASE

To the best of our knowledge, there is no commercially available multi-channel speaker recognition database for the TDSU task. Therefore, we designed our own database which consists of the recordings of 59 speakers over 5 different handset-channel conditions. In the database, there are 42 male and 17 female speakers. Each speaker repeats a single utterance, "benim parolam ses kaydımdır” (in English, “my password is my recording”), in which 5 of the 8 vowels in the Turkish language appear at least once. The recordings are taken in two separate sessions. In the first session, speakers repeat the utterance 5 times. In the second session 2 repetitions are recorded. Speakers are assisted with interactive voice response (IVR) prompts throughout the sessions. We record the utterances in different environments with varying background noise levels. However, we should mention that most of them are taken in a noisy office environment.

IVR system is implemented behind a public switched telephone network (PSTN) channel. We place calls to this system from five different handset-channel conditions:

1. A fixed wired analog phone, PSTN.
2. The same phone in the first condition but in hands-free mode, PSTN.
3. Another fixed wired analog phone, PSTN.
4. A fixed wireless digital phone, PSTN.
5. A fixed cell phone, GSM network.

Each condition in the above list is specifically chosen to represent distinct handset-channel conditions and background noise levels. First and third handsets are two different wired, analog phones. The second condition represents nosier environment compared to the others. However, it is more realistic for users who are often busy at work. In the fourth condition, we use a wireless telephone handset. A fixed cellular phone is used in the fifth condition. PSTN-PSTN connections are employed for the first four conditions in the database. GSM-PSTN channel is used in the fifth one.

3 METHODOLOGY

3.1 Baseline Spectral System

In the baseline spectral system, we first train a 64-state whole-phrase speaker independent HMM (SI HMM) for the fixed utterance using various recordings of the utterance from several different speakers and channels. The SI HMM has left-to-right topology without skip states. Each state has 4 mixtures. The model is adapted to speaker dependent models using MAP technique for only mean vector parameters. Time-normalized forced alignment likelihoods to the known text are used for decision making. Speaker likelihoods are obtained with the claimant speaker model and normalized with the SI HMM score.

In the sentence HMM method, varying length beginning and end silences might result in alignment problems and lead to significant degradation in verification accuracy. In order to avoid this problem, the silence sections are removed from each utterance prior to the feature extraction. We make use of speech recognition models for the silence removal. First, phone level alignment is performed for each utterance using tri-phone HMMs trained for the Turkish language using approximately 50 hours of speech. Then, silence aligned segments are clipped. The feature vectors are extracted from the clipped utterances. Each feature vector consists of 13 mel-frequency cepstral coefficients (MFCCs) (including zeroth value) and their first order derivatives. MFCCs are computed for 25 ms window length and 10 ms frame shift. They are normalized using cepstral mean normalization (CMN). The sentence HMM method is realized with HTK toolkit (Young et al., 2006).

3.2 Prosodic Systems

We need a time alignment technique in order to compare the prosodic features between enrollment and authentication utterances. For this purpose, we make use of the state alignment information in the sentence HMM method. In the procedure, each

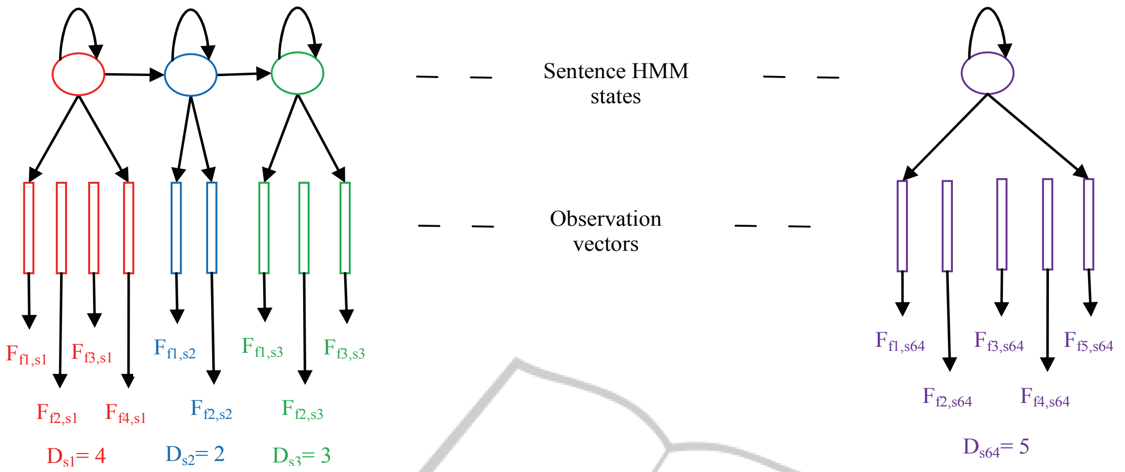

Figure 1: Alignment of the observation vectors to the sentence HMM states.

enrollment utterance is aligned using the speaker's own sentence HMM which is adapted with the same utterances. The claimant model is used for the alignment of the authentication utterances.

Alignment of the feature vectors to the sentence HMM states is illustrated in Figure 1. In the figure, pitch and energy features are represented with F symbol and duration feature is represented with D symbol. Fi,j is the pitch or log-energy value at frame i which is aligned to state j. Similarly, Dj is the duration value at state j. In our study, we use RAPT algorithm (Talkin, 1995) for pitch extraction.

During enrollment, we calculate pitch, energy and duration statistics for each sentence HMM state. Mean pitch (or energy) for an HMM state is calculated using the pitch (or energy) values of the frames which are aligned to the corresponding state as in Equation 1;

$$\mu_{F_j} = \frac{1}{A_j}\sum_{k=1}^{E}\sum_{i=1}^{A_j^k} F_{i,j}^k \qquad (1)$$

where the superscript k denotes the enrollment utterance number and E denotes the total number of enrollment utterances. Ajk is the number of frames aligned to state j in kth enrollment utterance and Aj is the total number of frames aligned to state j during the enrollment. Finally, μFj is mean pitch (or energy) for state j.

Mean duration for an HMM state is estimated using the number of frames aligned to the corresponding state as follows;

$$\mu_{D_j} = \frac{1}{E}\sum_{k=1}^{E}\frac{D_j^k}{N^k} \qquad (2)$$

where Nk is the total number of frames in kth enrollment utterance and μDj is mean duration for state j. As observed in Equation (2), duration feature is normalized with the total number of frames in the utterance.

During authentication, authentication scores for pitch-energy and duration systems are computed as in Equations (3) and (4), respectively;

$$\Gamma\_F = \frac{1}{A}\sum_{j=1}^{N}\sum_{i=1}^{A_j}\left|F_{i,j} - \mu_{F_j}\right| \qquad (3)$$

$$\Gamma\_D = \frac{1}{N}\sum_{j=1}^{N}\left|D_j - \mu_{D_j}\right| \qquad (4)$$

where N is the total number of states in the sentence HMM, A is the total number of frames in the utterance and Aj is the number of frames aligned to state j.

For authentication score calculation, we choose absolute difference after some informal trials. Also in (Charlet et al., 2000), the best performance is achieved with the same distance measure. In the prosodic systems, we do not make use of the variance parameter since we think that there is not adequate number of samples to reliably estimate the parameter. Our informal tests also verified this observation in terms of verification performance.

By deriving pitch and energy statistics from the state alignment information, we can trace and compare pitch and energy contours using acoustically relevant speech segments. Additionally, state durations might provide complementary information to the spectral system. In (Charlet et al., 2000), a simple integration of alignment and acoustic scores improved the verification accuracy.

3.3 System Fusion

The scores from the spectral and prosodic systems are combined with a three-layer perceptron network. The numbers of neurons in the layers are four, three and one, respectively. Transfer functions for the first two layers are hyperbolic tangent sigmoid. We use linear transfer function in the last layer.

4 EXPERIMENTS

4.1 Experimental Setup

Speakers in the database are divided into three categories for the experiments. These categories are named as background speakers, cohort speakers and test speakers. Ten speakers are set aside for SI HMM training in the spectral system. All utterances of the background speakers from five channel conditions are used to train the SI HMM. Forty speakers are used as cohorts to perform score normalization on the final likelihoods. Verification experiments are conducted with the remaining speakers' authentication utterances.

We prepare six different combinations of cohort and test speakers in order to increase the number of authentication trials. Five of the sets contain nine test speakers and the last set contains the remaining four speakers. Each speaker in the database is used only once as a test speaker and test and cohort speakers in the same set do not overlap with each other. For all test sets, the same background speakers are used. One of the sets is used to train the neural network parameters in Section 3.3., the other five sets are used in verification experiments.

In the spectral system, the SI HMM is adapted to speaker models using three utterances from the first session. The same adaptation procedure is employed for cohort and claimant models. Pitch, energy and duration statistics are extracted from the same three enrollment utterances. The remaining recordings of the test speakers are used in authentication. Each authentication utterance is used as a genuine trial for its own account and as imposter trials for the other speakers' account in the test set. The trials are performed for all possible enrollment-authentication channel combinations. For the test sets which contain nine speakers, 5 genuine (1 match + 4 mismatch condition) and 40 imposter (8 match + 32 mismatch condition) trials are carried out for each utterance. Total number of genuine and imposter trials in the five sets are 3885 and 29180, respectively.

4.2 Score Normalization

In order to compensate for the effects of channel mismatch conditions, we employ T-norm and HT-norm (Auckenthaler et al., 2000) score normalization techniques. In the techniques, each authentication utterance is scored against a set of example imposter models in parallel with the claimant model. Then, mean and standard deviation of the imposter scores are calculated. These parameters are used to perform the normalization in Equation (5).

$$NS = \frac{S - \mu_n}{\sigma_n} \qquad (5)$$

where µn and σn are the normalization parameters, S is the speaker score and NS is the normalized speaker score.

In T-norm, we do not assume any prior knowledge about the claimant's enrollment/authentication channel condition. Therefore, all five channel models of the cohorts are used in the normalization. In HT-norm, the parameters are estimated using the likelihoods of the cohorts who share the same channel type with the claimant's enrollment. In HT-norm, we assume that the claimant's enrollment channel is known and thus we benefit from this extra information to make better parameter estimation.

4.3 Verification Results

In Table 1, equal error rates (EERs) for the spectral and prosodic systems are presented for match and mismatch conditions. In the table, scores are normalized with T-norm. In Table 2, EERs for the spectral, prosodic and fusion systems are given for mixed condition in which match and mismatch trials are accumulated. In the fusion, we combine T-norm (or HT-norm) normalized scores of the spectral, duration and pitch systems. We did not use energy in the fusion since it did not provide any additional improvement.

We can make several observations from Tables 1 and 2. First, as observed in Table 1, channel mismatch conditions result in higher relative performance degradation in the spectral system when compared to the prosodic systems. Among the prosodic features, energy is more susceptible to mismatch conditions. This might be attributed to the differences in microphone and background noise levels in the recording sessions. Second, we observe that HT-norm improves the verification accuracy significantly in the spectral system while the rate of improvement is marginal in all prosodic systems.

Table 1: EERs for the spectral and prosodic systems in match and mismatch conditions.

	Match	Mismatch
Spectral (T-norm)	0.26	1.93
Duration (T-norm)	6.31	11.42
Pitch (T-norm)	13.51	16.41
Energy (T-norm)	11.84	30.44

Table 2: EERs for the spectral, prosodic and fusion systems in mixed (match + mismatch) condition.

	T-norm	HT-norm
Spectral	2.19	0.98
Duration	10.81	9.55
Pitch	15.98	14.98
Energy	28.93	25.92
Fusion (Spectral + Pitch + Duration)	1.96	0.88

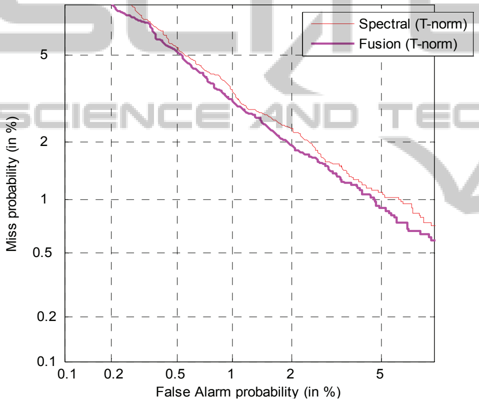

Figure 2: DET curves for the spectral and fusion systems. T-norm is used for score normalization.

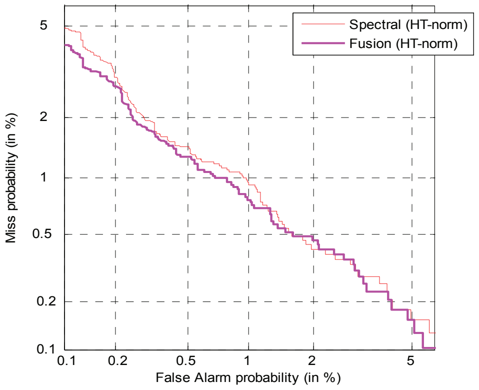

Figure 3: DET curves for the spectral and fusion systems. HT-norm is used for score normalization.

The first two observations indicate that prosodic features are more robust to channel variations. Additionally, we can conclude that providing handset-channel information is useful in score normalization process. Third, prosodic features do not yield high accuracy when they are individually employed. Among the prosodic features, the best performance is obtained for duration and the worst performance is obtained for energy. Fourth, the fusion of the spectral and prosodic systems improves the verification performance. Absolute reduction in EER is higher in T-norm when compared to HT-norm. On the other hand, relative reduction is approximately 10% for both normalization methods. Although prosodic features are found to be more robust to channel mismatch conditions, they do not provide higher relative improvement in handset-independent normalization. In Figure 2, detection error tradeoff (DET) curves for the spectral and fusion systems are depicted where the scores are normalized with T-norm. In Figure 3, DET curves for HT-norm scores are presented. As observed in the figures, the fusion outperforms the baseline spectral system in almost all operating points of the DET curves. These results show that prosodic features might provide complementary information to spectral features in a TDSU task.

5 CONCLUSIONS

Although high-level information sources have been extensively studied for text-independent tasks, less effort has been made to utilize them for text-dependent applications. In this study, we combined spectral and prosodic (pitch and duration) features together in order to improve the verification accuracy in a text-dependent single utterance speaker verification application. Recently, the target application has drawn more attention in private sector due to its ease of use and higher accuracy for relatively short utterances.

We made use of sentence HMM state alignment information to extract pitch and duration statistics. In our experiments, we observed that although the prosodic features individually do not yield high performance, they provide complementary information to the spectral features. We achieved approximately 10% relative reduction in EER when the scores from different sources are fused with a multi-layer neural network. Additionally, experimental results showed that prosodic features are more robust to channel mismatch conditions as expected.

All the experiments in this study are conducted using a relatively small database. In the future, we plan to collect larger databases for the TDSU task.

ACKNOWLEDGEMENTS

This work was supported in part by the Turkish State Planning Organization (DPT) under the TAM Project, number 2007K120610. This work was also supported in part by the TUBITAK TEYDEB 1509 under the Electronic Doctor's Round Project, number 9090036.

REFERENCES

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. "Score normalization for text-independent speaker verification systems," Digital Signal Processing 10 (1-3), pp. 42-54.

Charlet, D., Jouvet, D., Collin, O., 2000. "An alternative normalization scheme in HMM-based text-dependent speaker verification," Speech Communication 31 (2-3), pp. 113-120.

Dehak, N., Dumouchel, P., Kenny, P., 2007. "Modeling prosodic features with joint factor analysis for speaker verification," IEEE Transactions on Audio, Speech and Language Processing 15 (7), pp. 2095-2103.

Ferrer, L., Scheffer, N., Shriberg, E., 2010. "A comparison of approaches for modeling prosodic features in speaker recognition," International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010).

Klusacek, D., Navratil, J., Reynolds, D., Campbell, J., 2003. "Conditional pronunciation modeling in speaker detection," International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003).

NIST, 2012. "National Institute of Standards and Technology. Speaker Recognition Evaluation," http://www.nist.gov/speech/tests/spk.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang, B., 2003. "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003).

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., 2005. "Modeling prosodic feature sequences for speaker recognition," Speech Communication 46 (3-4), pp. 455-472.

Talkin, D., 1995. "A robust algorithm for pitch tracking (RAPT)", Speech Coding and Synthesis edited by W. B. Kleijn, K.K. Paliwal (Elsevier, New York), pp. 495–518.

Weber, F., Manganaro, L., Peskin, B., Shriberg, E., 2002. "Using prosodic and lexical information for speaker identification," International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002).

Yegnanarayana, B., Prasanna, S. R. M., Zachariah, J. M., Gupta, C.S., 2005. "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," IEEE Transactions on Speech and Audio Processing 13 (4), pp. 575-582.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department.