SPEAKER RECOGNITION IN NOISY CONDITIONS WITH LIMITED TRAINING DATA

Niall McLaughlin, Ji Ming, and Danny Crookes

Institute of ECIT, Queens University Belfast
Belfast, BT7 INN, UK
phone: +44 (0) 28 9097 1700, fax: +44 (0) 28 9097 1702, email: nmclaughlin02@qub.ac.uk
web: www.qub.ac.uk

ABSTRACT

In this paper we present a novel method for performing speaker recognition with very limited training data and in the presence of background noise. Similarity-based speaker recognition is considered so that speaker models can be created with limited training speech data. The proposed similarity is a form of cosine similarity used as a distance measure between speech feature vectors. Each speech frame is modelled using subband features, and into this framework, multicondition training and optimal feature selection are introduced, making the system capable of performing speaker recognition in the presence of realistic, time-varying noise, which is unknown during training. Speaker identification experiments were carried out using the SPIDRE database. The performance of the proposed new system for noise compensation is compared to that of an oracle model; the speaker identification accuracy for clean speech by the new system trained with limited training data is compared to that of a GMM trained with several minutes of speech. Both comparisons have demonstrated the effectiveness of the new model. Finally, experiments were carried out to test the new model for speaker identification given limited training data and with differing levels and types of realistic background noise. The results have demonstrated the robustness of the new system.

1. INTRODUCTION

Speaker identification becomes a difficult problem when the data used for identification is corrupted by background noise. This problem may be further compounded by a shortage of training data from each speaker. In this paper we consider the problem of speaker identification in a noise corrupted environment with limited training data i.e. there is insufficient data to build a statistical model e.g. a GMM for each speaker.

With limited training, a GMM speaker model may be obtained by adapting a universal background model (UBM) [1]. However, this approach is only possible if a UBM already exists. In this paper, we consider the applications without assuming the availability of an UBM for model adaptation, and assuming limited training speech for each speaker (e.g., of the scale of a few seconds). Recent work in the area of speaker recognition with limited training data has focused on building similarity, rather than probability, based speaker models. Fuzzy vector quantisation has been used to build speaker models that can be trained using only several seconds of data per speaker [2]. This approach can be extended by assigning different weights to different codewords representing a speaker, where the weights are dependent on each codeword's discriminative ability [3]. When speaker models are built with very limited training data, there may be significant overlap of data between the models. Removing the overlapping features has been shown to improve recognition performance by making each speaker model more discriminative [4]. Another non-statistical approach is to build speaker models using a linear classifier based on a variant of linear discriminant analysis (LDA). The work in [5] has attempted to overcome the problem of limited training data with LDA. In this paper, we present a new similarity-based approach to the limited training data problem. Our new similarity takes a form of a modified cosine similarity. Furthermore, we build a recognition system based on this new similarity which is capable of offering robustness to background noise with limited training data.

Speaker recognition becomes a much more difficult problem in the presence of environmental noise. This is because noise changes a speaker's acoustic features, making them different to those seen during training. Several approaches have been tried to improve the noise robustness of speaker recognition. With a priori knowledge of the noise characteristics it is possible to filter noise from speech, using techniques such as Kalman filtering [6] or spectral subtraction [7]. Instead of trying to remove background noise, it is possible to attempt to extract noise-robust features, e.g. RASTA features [8] from the speech signal. Missing feature theory may be used to ignore parts of speech corrupted by background noise [9] [10]. This may be extended using multicondition training to address full-band noise corruption [11]. Most of the above approaches have found use in statistical speaker models (e.g. GMMs or HMMs). In this paper, we consider noise compensation within our new similarity-based recognition system. We build on previous work, by combining missing feature theory with multicondition training. This allows our system to recognise a speaker in the presence of time-varying, unknown background noise with very limited speaker training data.

2. SIMILARITY-BASED SPEAKER RECOGNITION

We consider a short training speech segment for speaker λ of T frames $\mathbf{X}^{\lambda} = (\mathbf{x}^{\lambda}(1), \mathbf{x}^{\lambda}(2), ..., \mathbf{x}^{\lambda}(T))$, where each $\mathbf{x}^{\lambda}(t)$ is a frame vector at time t. We assume adverse test conditions in which the test speech may be corrupted by background noise. To accommodate the corruption, we represent each speech frame as F non-overlapped subbands, i.e., $\mathbf{x}^{\lambda}(t) = (x_1^{\lambda}(t), x_2^{\lambda}(t), ..., x_F^{\lambda}(t))$ where $x_f^{\lambda}(t)$ is the feature for subband f in frame $\mathbf{x}^{\lambda}(t)$. Improving robustness to corruption will be discussed in the next section. In this section, we focus on modelling the person given limited training data.

In recognition, let $\mathbf{Y} = (\mathbf{y}(1), \mathbf{y}(2), ..., \mathbf{y}(\Gamma))$ be a test

speech segment of Γ frames from an unknown speaker. Let $C(\mathbf{Y}, \mathbf{X}^\lambda)$ represent a similarity measure between the test sequence $\mathbf{Y}$ and a model sequence $\mathbf{X}^\lambda$. We identify the unknown person as follows, assuming text-independent training and test speech segments

$$\begin{aligned}\hat{\lambda} &= \arg\max_{\lambda} C(\mathbf{Y}, \mathbf{X}^\lambda) \\ &= \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau)) \end{aligned} \quad (1)$$

In (1), we assume that the overall similarity between the model and test sequences can be expressed as the sum of the similarities between their individual frames $C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau))$, where $\mathbf{x}^\lambda(\tau)$ is the model frame at time τ from person λ that matches test frame $\mathbf{y}(t)$. In order to perform text-independent speaker recognition, we select, for each test frame $\mathbf{y}(t)$, the best matching model frame $\mathbf{x}^\lambda(\tau)$ for comparison.

3. SIMILARITY MEASURE

We use a modified cosine similarity as the similarity measure $C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau))$ between speech frames. The standard cosine similarity is modified to be more robust to partial feature corruption. The cosine similarity $C(\mathbf{a}, \mathbf{b})$ between two vectors $\mathbf{a} = (a_1, a_2, ..., a_Q)$ and $\mathbf{b} = (b_1, b_2, ..., b_Q)$, each composed of Q local feature vectors, can be expressed as

$$\begin{aligned} C(\mathbf{a}, \mathbf{b}) &= \sum_{q=1}^{Q} \frac{a_q \cdot b_q}{||a_q|| ||b_q||} \frac{||a_q|| ||b_q||}{||\mathbf{a}|| ||\mathbf{b}||} \\ &= \sum_{q=1}^{Q} C(a_q, b_q) w_q \end{aligned} \quad (2)$$

where $C(a_q, b_q) = a_q \cdot b_q / ||a_q|| ||b_q||$ is the inner product between vectors a_q and b_q normalized by their respective norms. From (2) we can see that the overall cosine similarity is the sum of all the local cosine similarities $C(a_q, b_q)$ each weighted by w_q, which equals the norms of the appropriate local vectors compared to the norms of the overall vectors. As the weight w_q is a function of the overall norms, it will be affected by any local corruption in either $\mathbf{a}$ or $\mathbf{b}$. In other words, the weighting can spread local corruptions globally. To avoid this problem, we assume a uniform weight w_q for all the local vectors, meaning they contribute equally to the overall similarity. Thus, we use a uniformly-weighted cosine similarity as the measure $C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau))$ between two frames, defined in (1). This can be written as

$$C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau)) \simeq \sum_{f=1}^{F} C(y_f(t), x_f^\lambda(\tau))) \quad (3)$$

Note that the cosine similarity is bounded to between -1 and 1. This property will prove useful when we later express (3) in exponential form, as the value of the exponential will remain bounded, thus avoiding overflow problems.

4. ROBUSTNESS TO PARTIAL CORRUPTION

The system as currently defined, assumes that the test data is uncorrupted. We extend the system to be resistant to partially corrupted background noise by modifying the computation of the similarity measure $C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau))$, i.e. (3), between a noisy test frame $\mathbf{y}(t)$ and a clean model frame $\mathbf{x}^\lambda(\tau)$.

Assume that the given test frame $\mathbf{y}(t)$ contains noisy subbands, such that $\mathbf{y}(t)$ can be divided into two subsets $\mathbf{y}_\mathbf{f}(t)$ and its complement $\mathbf{y}_{\bar{\mathbf{f}}}(t)$. In this division, $\mathbf{y}_\mathbf{f}(t)$ is a feature set in the frame $\mathbf{y}(t)$ containing uncorrupted speech subbands, addressed by the subband index set $\mathbf{f} \subseteq [1, 2, ..., F]$. The complement $\mathbf{y}_{\bar{\mathbf{f}}}(t)$ contains the rest of the speech subbands which are considered unreliable due to corruption. Robustness to corruption can be achieved by replacing the full feature set with the reliable feature subset during the calculation of the similarities. Hence, the similarity between a noisy frame $\mathbf{y}(t)$ and a clean model frame $\mathbf{x}^\lambda(\tau)$ can be written as

$$\begin{aligned} C(\mathbf{y}(t), \mathbf{x}^\lambda(\tau)) &\simeq C(\mathbf{y}_\mathbf{f}(t), \mathbf{x}_\mathbf{f}^\lambda(\tau)) \\ &\simeq \sum_{f \in \mathbf{f}} C(y_f(t), x_f^\lambda(\tau)) \end{aligned} \quad (4)$$

Equation (4) is based on missing-feature theory or the "recognition by parts" principle [11][12]. The frame-by-frame processing indicated in (4) makes the system robust to time-varying noise.

In practical applications with unpredictable background noise, the clean-feature index set $\mathbf{f}$ is unknown *a priori*. In this paper, we develop an algorithm to obtain an estimate of $\mathbf{f}$ subject to an optimality criterion. The algorithm includes two steps. First, we express (4) in an equivalent exponential form (in terms of recognition based on maximum similarity)

$$\begin{aligned} p(\mathbf{y}_\mathbf{f}(t) | \mathbf{x}_\mathbf{f}^\lambda(\tau)) &= H^{C(\mathbf{y}_\mathbf{f}(t), \mathbf{x}_\mathbf{f}^\lambda(\tau))} \\ &= \prod_{f \in \mathbf{f}} H^{C(y_f(t), x_f^\lambda(\tau))} \end{aligned} \quad (5)$$

where $H > 1$ is a positive base number. The function $p(\mathbf{y}_\mathbf{f}(t) | \mathbf{x}_\mathbf{f}^\lambda(\tau))$ shares the characteristics of an exponent-type likelihood function for the test feature set $\mathbf{y}_\mathbf{f}(t)$ associated with speaker λ, given the model feature set $\mathbf{x}_\mathbf{f}^\lambda(\tau)$, with unknown optimal feature index set $\mathbf{f}$. Second, based on this likelihood function, we create a 'posterior' probability of the model feature set, as a function of the optimal feature indexes. This can be written as

$$P(\mathbf{x}_\mathbf{f}^\lambda(\tau) | \mathbf{y}_\mathbf{f}(t)) = \frac{p(\mathbf{y}_\mathbf{f}(t) | \mathbf{x}_\mathbf{f}^\lambda(\tau))}{\sum_{\lambda'} \sum_{\mathbf{z}_\mathbf{f}^{\lambda'}} p(\mathbf{y}_\mathbf{f}(t) | \mathbf{z}_\mathbf{f}^{\lambda'}) + \varepsilon} \quad (6)$$

where we assume an equal prior probability for all the speakers. The sum in the denominator is over all model feature sets of all the speakers that may match the given test feature set, and ε is a small positive number accounting for any test set $\mathbf{y}_\mathbf{f}(t)$ without matching model set $\mathbf{x}_\mathbf{f}(\tau)$ (hence the sum approaches zero). Based on (6), the recognition rule (1) can be modified as follows to accommodate partial speech corruption.

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} \max_{\mathbf{f}} \log P(\mathbf{x}_\mathbf{f}^\lambda(\tau) | \mathbf{y}_\mathbf{f}(t)) \quad (7)$$

This expression seeks to find the most-likely speaker by jointly maximizing the similarity of the test sequence over all persons and all possible local feature sets.

5. ROBUSTNESS TO FULL-BAND SPEECH CORRUPTION

So far we have defined a system that achieves robustness to partial feature corruption by finding an optimal estimate of the uncorrupted feature set. To accommodate full feature set corruption, we extend the system by combining multicondition model training with optimal feature estimation. In this new framework, multicondition model training is used to provide coarse-grained compensation for the effect of unknown corruption. Optimal feature estimation is then used to refine this compensation by ignoring the remaining mismatched local features. These two steps combined offer robustness to full-set feature corruption.

Let $\mathbf{X}^{\lambda} = (\mathbf{x}^{\lambda}(1),..,\mathbf{x}^{\lambda}(T))$ be the given speech training sequence for speaker λ, and $\mathbf{X}^{\lambda,l} = (\mathbf{x}^{\lambda,l}(1),..,\mathbf{x}^{\lambda,l}(T))$, $l = 0,1,...,L$ represent $L+1$ multicondition training sequences generated from $\mathbf{X}^{\lambda}$, where each $\mathbf{X}^{\lambda,l}$ simulates a different corruption condition, with $\mathbf{X}^{\lambda,0}$ corresponding to the clean condition. These multicondition training sequences are combined to model test speech sequence $\mathbf{Y} = (\mathbf{y}(1),\mathbf{y}(2),...,\mathbf{y}(\Gamma))$ with unknown full feature set corruption. Previously, to compensate for partial corruption, we divided each test frame $\mathbf{y}(t)$ into two subsets, containing the corrupted and uncorrupted local feature sets, with respect to the clean training data. Now, to compensate for full feature set corruption we form $L+1$ subsets of $\mathbf{y}(t)$, each subset containing the features matched by a corresponding training condition from the $L+1$ multicondition training data.

Let $p(\mathbf{y}_{\mathbf{f}}(t)|\mathbf{x}_{\mathbf{f}}^{\lambda,l}(\tau))$ represent the likelihood of the test frame $\mathbf{y}(t)$ associated with a model frame $\mathbf{x}^{\lambda,l}(\tau)$ corrupted at training condition l, with index set $\mathbf{f}$ identifying their matching local features. We calculate $p(\mathbf{y}_{\mathbf{f}}(t)|\mathbf{x}_{\mathbf{f}}^{\lambda,l}(\tau))$ by replacing in (5) the model frame $\mathbf{x}_{\mathbf{f}}^{\lambda}(\tau)$ with the corresponding corrupted model frame $\mathbf{x}_{\mathbf{f}}^{\lambda,l}(\tau)$. Based on this likelihood function, we can create a 'posterior' probability of the model feature set $\mathbf{x}_{\mathbf{f}}^{\lambda,l}(\tau)$ at condition l, as a function of the optimal feature subset $\mathbf{f}$. This can be written as

$$P(\mathbf{x}_{\mathbf{f}}^{\lambda,l}(\tau)|\mathbf{y}_{\mathbf{f}}(t)) = \frac{p(\mathbf{y}_{\mathbf{f}}(t)|\mathbf{x}_{\mathbf{f}}^{\lambda,l}(\tau))}{\sum_{l'=0}^{L}\sum_{\lambda'}\sum_{\mathbf{z}_{\mathbf{f}}^{\lambda',l'}} p(\mathbf{y}_{\mathbf{f}}(t)|\mathbf{z}_{\mathbf{f}}^{\lambda',l'}) + \varepsilon} \quad (8)$$

The sum in the denominator is over all model feature sets of all the speakers, taking into account all the corrupted training conditions (including the clean training condition). As in the previous model (6), we obtain an optimal estimate of the unknown feature index set $\mathbf{f}$ at each corruption condition, by maximising the appropriate posterior probability (8). Therefore, the new recognition rule, which combines both multicondition training and optimal feature estimation, can be expressed as

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} \log\Big\{ \sum_{l=0}^{L} \max_{\mathbf{f}} P(\mathbf{x}_{\mathbf{f}}^{\lambda,l}(\tau)|\mathbf{y}_{\mathbf{f}}(t))\Big\} \quad (9)$$

In (9) the optimal feature set is estimated at each training condition, and contributions of all the training conditions are summed towards the overall similarity of each frame. The above defined system offers robustness to full-band corruption without assuming knowledge of noise, and can be trained with only limited data from each speaker.

6. EXPERIMENTS

In this section we test the performance of the proposed system for speaker identification with limited training data and in the presence of background noise. To the best of our knowledge, the use of similarity-based speaker recognition with modified cosine-similarity, together with optimal feature selection and multicondition training appears to be unique in the literature. We perform a speaker identification experiment using the SPIDRE [13] database. Performance in this experiment was compared to the published results of an of an existing GMM system tested with the same database. Speaker identification performance with very limited training data (as little as 3 s of speech) was compared to the published results from an existing system [2] designed for this purpose, using the YOHO speaker database [14].

The noise compensation performance of the system has been tested against an oracle model using speech samples corrupted with band-limited noise. The oracle model used prior knowledge of the noise location to remove the corrupted subbands, while the test system, without knowledge of the noise location or characteristics, selected the optimal subbands as discussed in the above algorithms. Performance of the system in the presence of real-world background noise was tested by adding realistic, full-band, time-varying, noise to test samples from the SPIDRE database at a variety of signal to noise ratios (SNRs).

6.1 Speaker identification with limited training data

We evaluate the performance of our system by comparing identification accuracy with limited training to that of published results produced using two different databases.

An experiment was carried out using the SPIDRE database. The SPIDRE database, a subset of the Switchboard database, consists of four conversation halves from 45 speakers (27 male, 18 female). We use the A1 and A2 conversation halves for testing and training respectively. Our system trained with limited data was compared against a GMM based system trained with several minutes of data per speaker. Each speech sample was silence-stripped and divided into 20 ms frames overlapping by 10 ms. Each frame was processed through a 22-channel log mel-scale filter and the filter outputs decorrelated with a high pass filter, giving 21 decorrelated log mel filter bank coefficients. These coefficients were uniformly placed into groups of three, giving seven subband features. First-order derivative coefficients were included, resulting in 14-subband feature streams for each frame, each stream containing three elements. Speaker models were constructed using segments of speech from each speaker of varying durations from 2 s to 30 s. Speaker identification experiments were performed using three 10 s samples of testing data from each speaker.

Table 1 presents speaker identification accuracy as a function of the training data duration and test data duration. The results from our system are compared with a previously published result based on a 32-mixture GMM for each speaker trained using all the available training data (about two minutes per speaker) [15]. From these results we can see that identification accuracy with 30 s of training data is comparable with the GMM-based result. In fact when our new system is trained with 30 s of speech and tested with 10 s of speech its identification accuracy exceeds that of the GMM system. In addition, we can see that testing with a larger

Table 1: *Speaker identification accuracy (%) for 5 s and 10 s testing data on the SPIDRE database with varying training duration. This table compares our new method of similarity-based speaker identification against a GMM trained with two minutes of data.*

	Similarity-based speaker recognition					GMM
Train/ Test	30 s	20 s	10 s	5 s	2 s	∼2min
10 s	93.3	85.9	81.4	77.7	62.9	89.9
5 s	87.4	85.2	80	74.8	57.7	89.9

amount of data (moving from 5 s testing to 10 s testing) improves identification accuracy. The results indicate that our recognition system based on modified cosine similarity is a viable method for speaker identification.

In order to test the speaker identification performance of our proposed system with extremely limited training data, we replicate the testing conditions in [2]. This published system performed speaker identification using only a few seconds of training and testing data per-speaker, by using fuzzy vector-quantization speaker models. The first 30 speakers from the YOHO database [14] were used in this experiment. We used the same acoustic features as above for modelling the speech. In common with [2] we used the same length of testing and training data for each experiment (i.e. in the 3 s experiment, we used 3 s of training data and 3 s testing data per speaker). Tests were performed with 3, 6 and 12 s of testing and training data.

Table 2: *Speaker identification accuracy (%) tested on the YOHO speaker database, compared to published results from [2] with varying training duration. (Note, the figures for 6s and 12s quoted from [2] are estimated from Fig.2 in that publication, as the numerical values are not given).*

Training duration (s)	3	6	12
Our system	84.4	91	100
(MFCC+Δ+ΔΔ+LP-residual)	86.7	94	97
(MFCC+Δ+ΔΔ)	80	83	94

The results from this experiment are presented in Table 2. With 3 s of testing and training data our system has a speaker identification accuracy of 84.4%, this improves over the accuracy produced using fuzzy vector quantization (FVQ) with MFCC+Δ+ΔΔ features of 80%. The accuracy of our system is almost as good as the best result produced using FVQ with MFCC+Δ+ΔΔ+LP-residual features of 86.67%. From examining the results of [2] we can say that the feature-type used has a large impact on identification accuracy. It may be the case that our system would show improved performance if used with different features. Table 2 shows further results for 6 s and 12 s of testing and training data. Again we can see that the results produced by our system are comparable with those obtained using FVQ together with MFCC+Δ+ΔΔ+LP-residual features.

6.2 Speaker identification with unknown noise corruption compared to an oracle model

For this experiment test samples from each speaker were corrupted using band-limited noise at various different centre frequencies and bandwidths. The oracle model used prior knowledge of the corruption to remove the affected subbands before performing recognition, hence performing as an idealised, noise-robust system might be expected to. Our system did not have knowledge of the characteristics of the noise corruption and performed recognition by optimally selecting the subbands. In addition, a baseline system which performed recognition using all the subbands was also tested; this is referred to as the 'do-nothing' system.

Noise corrupted test samples were generated by adding 0 dB SNR band-limited noise to clean speech test samples from the SPIDRE database. In this experiment the system was trained using 30 s of clean speech and tested using five 10 s samples. In each test sample, different band-limited noises were created to corrupt different numbers of adjacent subbands. The results of this experiment are show in Table 3, as a function of the number of affected subbands.

Table 3: *Speaker identification accuracy (%) with variable numbers of subbands corrupted, comparing our system to the oracle model and 'do nothing' baseline. The Centre column gives the centre frequency and the B/width column gives the bandwidth of the corrupting noise.*

Corruption Properties			Accuracy (%)		
Centre (Hz)	B/width (Hz)	Noisy Bands	Our System	Oracle	Do Nothing
656	175	1	68	65.3	58.7
1031	225	2	64	60	51.6
1265	325	3	46.2	45.3	28.9
2156	400	3	47.6	48	28.9

From the results in Table 3 we see that for the most part our proposed system performed better than the oracle model and always significantly better than the 'do-nothing' model. This indicates that the system is capable of removing the contribution of noise corrupted speech subbands from each speaker's score. The fact that our system could outperform the oracle model in many cases, may be due to the fact the oracle model removes all bands believed to be corrupted by noise. It may be the case that some features e.g. the delta features of some corrupted bands, are only partially corrupt and thus are still usable for recognition. This can occur because the added noise corruption does not have a sharp cut-off frequency. Our system may be taking advantage of this information, ignored by the oracle model, to produce more accurate identification scores.

6.3 Speaker identification with more realistic noise corruption

A speaker identification experiment was performed with noise corrupted test samples, created by adding realistic non-stationary, full-band noise at 10 dB and 15 dB SNR to each test sample from the SPIDRE database. The background noise types used were pop-song, restaurant, and street noise. Noisy speaker models were created using multicondition training, by adding low-pass filtered white noise, with a 3-dB cutoff frequency of 2 kHz, at SNRs from 10 dB to 20 dB in 5 dB steps to each clean training segment. Tests were carried out using three 10 s testing samples for each person. In order to assess the contribution of multicondition training and optimal feature selection to the overall score several

Table 4: *Speaker identification accuracy (%) tested on the SPIDRE speaker database with various realistic noise types added at variable SNRs. Tests were performed with both 5 s and 10 s of training data per speaker. These results were produced with three 10 s samples testing data per speaker.*

		Noise Condition						
		Clean	Restaurant		Street		Pop-song	
System	Training (s)		10dB	15dB	10dB	15dB	10dB	15dB
Our System	5	79.3	54.1	67.4	59.3	73.3	66.7	74.1
	10	81.5	63.7	73.3	72.6	76.3	77	78.5
Optimal Feature Only	5	82.2	39.3	55.6	41.5	55.6	59.3	71.1
	10	82.2	45.1	62.2	42.9	62.2	68.8	76.3
Multicondition Only	5	72.6	41.5	51.8	55.5	68.1	60	67.4
	10	81.5	51.9	66.7	65.9	74.8	69.6	79.3
Do Nothing	5	80	31.9	47.7	39.3	51.9	56.3	68.8
	10	83.7	42.2	57.7	42.2	61.5	65.9	76.2

versions of the system were tested. In addition to our main system which combines optimal feature selection with multicondition training, we performed tests with optimal feature selection only, multicondition training only, and neither technique, the last corresponding to a 'do-nothing' system.

The results of these experiments are presented in Table 4. It is apparent from these results that our new system improved over the do-nothing system in all the noisy test conditions, and experienced a slight performance loss in the clean condition. The table also showed the independent contributions of the multicondition training and optimal feature selection in the new system. In general, with only one exception where multicondition training outperformed our new system, each technique offered an improvement over the do-nothing model, and the combination of these techniques, in our new system, resulted in greater improvement.

7. CONCLUSIONS

In this paper we have proposed a novel method of similarity-based speaker identification that can be used with very limited training data and in the presence of unknown background noise. We used a modified cosine similarity measure to compare speech features, and used multicondition training with optimal band selection to accommodate unknown noise. For clean data speaker recognition, experiments on two different databases showed that the new system achieved comparable performance with baselines. For speaker recognition using noisy test data and given very limited training data, which is a relatively new research area, our new system has shown significantly improved robustness over the baseline systems.

REFERENCES

[1] P. Angkititrakul and J. H. L. Hansen, "Discriminative in-set/out-of-set speaker recognition," *Audio, Speech, and Language Processing*, vol. 15, pp. 498, 2007.

[2] H. S. Jayanna and S. R. Mahadeva Prasanna, "Fuzzy vector quantization for speaker recognition under limited data conditions," *TENCON 2008*, pp. 1–4, 2009.

[3] C. Jian L. Lin and S. Xiaoying, "A discriminative method for speaker identification with limited data," *(FSKD)*, vol. 2, pp. 512, 2010.

[4] S. Kwon and S. Narayanan, "Robust speaker identification based on selective use of feature vectors," *Pattern Recognition Letters*, vol. 28, pp. 85–89, 2007.

[5] A.E. Rosenberg L. Qi, S. Parthasarathy and D.W. Tufts, "Normalized discriminant analysis with application to a hybrid speaker-verification system," *ICASSP-96*, vol. 2, pp. 681, 1996.

[6] T. Fingscheidt C. Beaugeant Suhadi, S. Stan, "An evaluation of vts and imm for speaker verification in noise," *EUROSPEECH-2003*, pp. 1669–1672, 2003.

[7] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," *ICSLP 96*, vol. 2, pp. 929–932, 1996.

[8] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," *ICASSP-98*, vol. 1, pp. 121–124, 1998.

[10] J.F. Bonastre L. Besacier and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Commun.*, vol. 31, pp. 89–106, 2000.

[11] J.R. Glass J. Ming, T.J. Hazen and D.A. Reynolds, "Robust speaker recognition in noisy conditions," *ASLP*, vol. 15, pp. 1711, 2007.

[12] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 101–116, 2005.

[13] D.A. Reynolds, "The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus," *ICASSP-96*, vol. 1, pp. 113, 1996.

[14] J.P. Campbell, "Testing with the yoho cd-rom voice verification corpus," *ICASSP-95*, pp. 341–344, 1995.

[15] D. Stewart J. Ming and S. Vaseghi, "Speaker identification in unknown noisy conditions - a universal compensation approach," *ICASSP '05*, vol. 1, pp. 617, 2005.