Speaker recognition via fusion of subglottal features and MFCCs

Harish Arsikere[1], *Hitesh Anand Gupta*[1], *Abeer Alwan*[1]

[1]Electrical Engineering Department, University of California, Los Angeles, CA 90095, USA

harishan@ucla.edu, hiteshag@ucla.edu, alwan@seas.ucla.edu

Abstract

Motivated by the speaker-specificity and stationarity of subglottal acoustics, this paper investigates the utility of subglottal cepstral coefficients (SGCCs) for speaker identification (SID) and verification (SV). SGCCs can be computed using accelerometer recordings of subglottal acoustics, but such an approach is infeasible in real-world scenarios. To estimate SGCCs from speech signals, we adopt the Bayesian minimum mean squared error (MMSE) estimator proposed in the speech-to-articulatory inversion literature. The joint distribution of SGCCs and speech MFCCs is modeled using the WashU-UCLA corpus (containing simultaneous recordings of speech and subglottal acoustics), and the resulting model is used to obtain an MMSE estimate of SGCCs from unseen (test) MFCCs. Cross-validation experiments on the WashU-UCLA corpus show that the estimation efficacy, on average, is speaker dependent. A score-level fusion of MFCC and SGCC systems outperforms the MFCC-only baseline in both SID and SV tasks. On the TIMIT database (SID), the relative reduction in identification error is 16, 40 and 51% for G.712-filtered (300–3400 Hz), narrowband (0–4000 Hz) and wideband (0–8000 Hz) speech, respectively. On the NIST 2008 database (SV), the relative reduction in equal error rate is 4 and 11% for 10 and 5 second utterances, respectively.

Index Terms: speaker recognition, subglottal acoustics, cepstral coefficients, score combination, MMSE estimation

1. Introduction

Speaker identification (SID) and verification (SV) are closely-related problems; they are jointly referred to as *speaker recognition*. Mel-frequency cepstral coefficients (MFCCs), which capture the acoustics of the supraglottal vocal tract, have been widely used for both tasks. They have been shown to provide good performance with a number of modeling schemes such as simple Gaussian mixture models (GMMs) [1], GMMs adapted from universal background models (UBMs) [2], support vector machine (SVM) supervectors [3, 4], joint speaker and channel factors [5, 6], and total-variability *i*-vectors [7]. Other features that have been proposed and used in conjunction with MFCCs (via feature-level or score-level fusion) include those based on voice-source parameters [8], spectro-temporal modulation frequencies [9], prosody [10], word patterns and lexicon [11], and articulatory parameters [12]. In this paper, we investigate the utility of *subglottal* features (capturing the acoustics of the tracheo-bronchial airways) for both SID and SV. The focus is specifically on cepstral coefficients extracted from subglottal acoustics (henceforth referred to as SGCCs) and their fusion with MFCCs for improved speaker-recognition performance.

To record subglottal acoustics, a noninvasive device called the accelerometer is generally used. In the past, we have studied the properties of subglottal resonances (SGRs)—by manually analyzing accelerometer recordings—and also developed automatic algorithms to estimate SGRs from speech signals [13]. Although we found speech-based SGR estimates to be effective for speaker height estimation and adaptation (especially in limited-data conditions) [13, 14, 15], pilot experiments on the TIMIT database showed that they were not discriminative enough for speaker recognition. Therefore, in this paper, we employ more informative spectral features in the form of SGCCs—they are computed just like MFCCs, except that they are based on subglottal acoustics instead of speech.

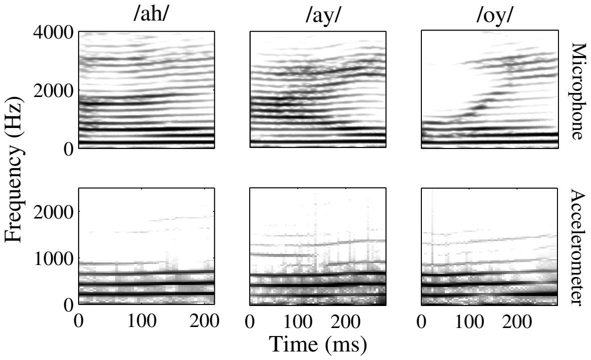

Figure 1: Vowel spectrograms comparing the within-speaker variability of speech (top panel) and subglottal acoustics (bottom panel). Data are sampled from the recordings of a female speaker in the WashU-UCLA corpus.

We are interested in subglottal features for two reasons. First, subglottal acoustics are speaker specific to some extent owing to their dependence on body height [16]. Second, the spectral characteristics of subglottal acoustics (for a given speaker) are much less variable than the spectral characteristics of speech. Figure 1 exemplifies this using vowel spectrograms of speech and their corresponding recordings of subglottal acoustics (data were obtained from the WashU-UCLA corpus [17]). The stationary nature of subglottal acoustics can be particularly beneficial when the amount of speech data (for enrollment and/or evaluation) is limited. One of the challenges, however, is to be able to estimate subglottal features (SGCCs) using speech, thus obviating the need for an accelerometer in real-world scenarios.

Our approach to estimating SGCCs from MFCCs is inspired by previous studies on speech-to-articulatory inversion [18, 19, 12]. The method proposed in these studies was to train joint statistical models from simultaneously-recorded speech and articulatory data, and then use those models to estimate articulatory trajectories from unseen utterances. Our approach to SGCC estimation is similar, except that we use simultaneous recordings of speech and subglottal acoustics.

In [12], articulatory parameters (estimated from speech sig-

Work supported in part by NSF Grant No. 0905381.

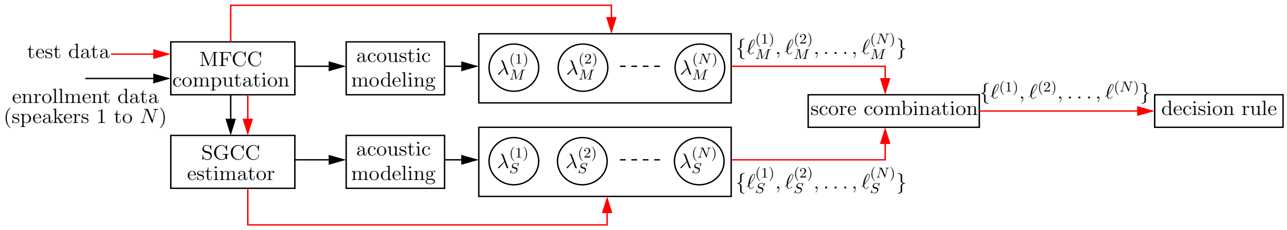
test data
enrollment data
(speakers 1 to N)
MFCC computation
SGCC estimator
acoustic modeling
acoustic modeling
$\lambda_M^{(1)}$ $\lambda_M^{(2)}$ ---- $\lambda_M^{(N)}$
$\lambda_S^{(1)}$ $\lambda_S^{(2)}$ ---- $\lambda_S^{(N)}$
$\{\ell_M^{(1)}, \ell_M^{(2)}, \ldots, \ell_M^{(N)}\}$
$\{\ell_S^{(1)}, \ell_S^{(2)}, \ldots, \ell_S^{(N)}\}$
score combination
$\{\ell^{(1)}, \ell^{(2)}, \ldots, \ell^{(N)}\}$
decision rule

Figure 2: Block diagram of the proposed SID/SV framework. The arrows in black correspond to training (enrollment) and the arrows in red correspond to evaluation. Subscripts M and S denote MFCCs and SGCCs, respectively. The λs denote speaker models and the ℓs denote acoustic model scores (for test data).

nals) were combined with MFCCs in an SV task. Using the classical UBM-GMM setup, it was shown that the combined system improved verification performance by 9–14% relative to the MFCC-only baseline. This study uses a feature-combination approach like [12], but with two important differences. (1) The production-based features used here (SGCCs) are those of the subglottal, not supraglottal, system. (2) In [12], a subset of the Wisconsin X-Ray Microbeam (XRMB) database (46 speakers) was used for SV experiments. Here, we evaluate our approach on databases that are larger and also more commonly used for speaker recognition. Using the TIMIT database for SID and the NIST 2008 database for SV, we show that SGCCs offer complementary information to the MFCC-only system.

2. Proposed framework

We propose a score-level framework to fuse the information provided by MFCCs and SGCCs (it will be explained later why feature concatenation is difficult). An overview of the proposed framework is presented here (see Figure 2) and the implementation details are provided in Section 4.

Let the number of speakers to be enrolled for SID or SV be N. Enrollment data are used to train two sets of acoustic models: $\{\lambda_M^{(1)}, ..., \lambda_M^{(N)}\}$ for MFCCs, and $\{\lambda_S^{(1)}, ..., \lambda_S^{(N)}\}$ for estimated SGCCs (details about the SGCC estimator are provided in Section 3). Given an unseen test utterance, MFCC and SGCC scores ($\{\ell_M^{(1)}, ..., \ell_M^{(N)}\}$, $\{\ell_S^{(1)}, ..., \ell_S^{(N)}\}$) are computed with respect to the pre-trained models and then combined in a weighted fashion. The combined scores $\{\ell^{(1)}, ..., \ell^{(N)}\}$ are used to make a decision (binary for SV, and 1-of-N for SID).

3. Estimating SGCCs using MFCCs

In [18], a Bayesian minimum mean squared error (MMSE) estimator was proposed for estimating articulatory parameters from speech acoustics. We adopt that approach here for SGCC estimation and evaluate it using the WashU-UCLA corpus (which contains time-synchronized recordings of speech and subglottal acoustics). The basic mathematical framework for MMSE estimation is provided below (see [18] for a detailed derivation) and the implementation details are deferred to Section 3.1.

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M]^\top$ and $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_S]^\top$ be M- and S-dimensional random vectors denoting MFCCs and the corresponding time-synchronized SGCCs, respectively. Let $\mathbf{Z} = [\mathbf{X}^\top \mathbf{Y}^\top]^\top$ denote the joint random vector. Since the distribution of $\mathbf{Z}$ is usually unknown, the simplest way to model it would be via a K-component GMM $\lambda^{(\mathbf{Z})}$:

$$p(\mathbf{z}|\lambda^{(\mathbf{Z})}) = \sum_{k=1}^{K} \nu_k^{(\mathbf{Z})} \mathcal{N}(\mathbf{z}; \mu_k^{(\mathbf{Z})}, \Sigma_k^{(\mathbf{Z})}), \tag{1}$$

where $\nu_k \mathcal{N}(\cdot; \mu_k, \Sigma_k)$ denotes the probability density function of the k^{th} mixture component, with mean μ_k, covariance Σ_k and weight ν_k. Once $\lambda^{(\mathbf{Z})}$ is available (from joint training data), the marginal and joint statistics of $\mathbf{X}$ and $\mathbf{Y}$ can be obtained using Eq. (2). Note that $\mathbf{Z}$ must be modeled using full covariances in order to extract the joint statistics of $\mathbf{X}$ and $\mathbf{Y}$.

$$\mu_k^{(\mathbf{Z})} = \begin{bmatrix} \mu_k^{(\mathbf{X})} \\ \mu_k^{(\mathbf{Y})} \end{bmatrix}, \quad \Sigma_k^{(\mathbf{Z})} = \begin{bmatrix} \Sigma_k^{(\mathbf{XX})} & \Sigma_k^{(\mathbf{XY})} \\ \Sigma_k^{(\mathbf{YX})} & \Sigma_k^{(\mathbf{YY})} \end{bmatrix} \tag{2}$$

Given an unseen test utterance, a sequence of MFCC vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ is first extracted from it. Then, for a given MFCC vector $\mathbf{x}_t$ ($1 \leq t \leq T$), the SGCC vector is computed as the conditional mean (or MMSE estimate) of $\mathbf{Y}$:

$$\hat{\mathbf{y}}_t = E[\mathbf{Y}|\mathbf{x}_t] = \sum_{k=1}^{K} P(k|\mathbf{x}_t, \lambda^{(\mathbf{Z})}) \zeta_{k,t}^{(\mathbf{Y})}, \tag{3}$$

where $E[\cdot]$ denotes the expectation operator, and $P(k|\mathbf{x}_t, \lambda^{(\mathbf{Z})})$ and $\zeta_{k,t}^{(\mathbf{Y})}$ are defined as in Eqs. (4) and (5), respectively.

$$P(k|\mathbf{x}_t, \lambda^{(\mathbf{Z})}) = \frac{\nu_k^{(\mathbf{Z})} \mathcal{N}(\mathbf{x}_t; \mu_k^{(\mathbf{X})}, \Sigma_k^{(\mathbf{XX})})}{\sum_{k'=1}^{K} \nu_{k'}^{(\mathbf{Z})} \mathcal{N}(\mathbf{x}_t; \mu_{k'}^{(\mathbf{X})}, \Sigma_{k'}^{(\mathbf{XX})})} \tag{4}$$

$$\zeta_{k,t}^{(\mathbf{Y})} = \mu_k^{(\mathbf{Y})} + \Sigma_k^{(\mathbf{YX})} {\Sigma_k^{(\mathbf{XX})}}^{-1} (\mathbf{x}_t - \mu_k^{(\mathbf{X})}) \tag{5}$$

In the present study, the MMSE estimator of Eq. (3) provides a mapping from the more-variable MFCC space to the less-variable SGCC space (can be viewed in some sense as a many-to-one mapping). On the other hand, in [18], the same MMSE estimator provides a one-to-many mapping from speech acoustics to articulatory parameters.

3.1. Implementation details and evaluation setup

The databases used by studies on speech-to-articulatory inversion consist of read speech utterances (and time-synchronized articulatory trajectories) with good phonetic and lexical coverage. The WashU-UCLA corpus, in contrast, consists only of short phrases of the form “I said a h[V]d again,” where [V] is one of 9 monophthongs, 4 diphthongs, or [r] (the corpus has 10 repetitions of each phrase from 50 adult speakers of American English—25 male and 25 female). To avoid redundancy in the training data that are used to estimate $\lambda^{(\mathbf{Z})}$ (note that all phrases have the same content except for the vowel [V]), only the vowel segments are isolated and used. However, since vowels form only a part of the speakers' phonetic space, we need a way to deal with non-vowel segments while estimating SGCCs for speaker recognition. Section 4 explains how this is done.

The MMSE estimator (described in Section 3) is evaluated using 5-fold cross validation. The available vowel samples (7000 in total: 50 speakers, 14 vowels, 10 repetitions) are split into 5 sets such that the data from any given speaker belong to exactly one set. All signals are down sampled to 8 kHz (from their original sampling rate of 48 kHz). MFCCs and SGCCs are extracted at 5 ms intervals using a 20 ms Hamming window and a 26-channel Mel filter bank. The zeroth cepstral coefficient is

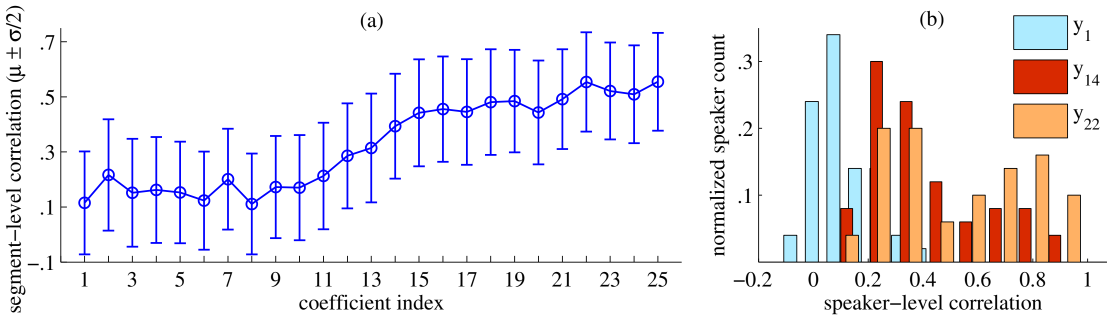

Figure 3: (a) Means (circles) and standard deviations (error bars) of the segment-level correlations (segment = vowel token) between actual and estimated SGCCs. Results from all 50 speakers in the WashU-UCLA corpus are pooled together. (b) Distribution of speaker-level correlation (i.e., average segment-level correlation on a per-speaker basis) for three different cepstral coefficients (y_1, y_{14}, y_{22}).

Feature set	J-Ratio
MFCCs (x_1–x_{25})	5.32
Actual SGCCs (y_1–y_{25})	5.89
Estimated SGCCs	5.79
MFCCs + actual SGCCs	9.04
MFCCs + estimated SGCCs	8.79

Table 1: J-ratio, a measure of class separation (class = speaker), for different features (+ denotes concatenation). Features were extracted from isolated vowel recordings of speech and subglottal acoustics, for all 50 speakers in the WashU-UCLA corpus.

discarded; MFCCs x_1–x_{25} and SGCCs y_1–y_{25} are used to train $\lambda^{(\mathbf{Z})}$. The number of components K is set to 16—roughly one component per vowel (we did not observe any significant improvements in performance by increasing K beyond 16).

3.2. Results

SGCC estimates from all 5 test sets are pooled together for analysis. The utility of the estimates (for speaker recognition) is assessed in two ways: (1) by computing the correlation between actual and estimated SGCCs on a per-segment basis (i.e., correlations between actual and estimated time trajectories), and (2) by comparing actual and estimated SGCCs with regard to their ability to discriminate between speakers.

Figure 3(a) shows the average segment-level correlations (with error bars) for SGCCs y_1 to y_{25}—the values lie between 0.12 and 0.55, and are comparable to the correlations achieved for speech-to-articulatory inversion [19]. An important observation from Figure 3(a) is that the error bars are significantly large, suggesting a high degree of speaker variability in the estimator's performance. Figure 3(b) verifies this further via distributions of the speaker-level correlation (i.e., average segment-level correlation on a per-speaker basis). In essence, we can attribute the discriminatory power of estimated SGCCs, in part, to the speaker-dependent nature of the MMSE estimator.

We use the J-Ratio [20], a popular measure of class separation, to compare the actual and estimated SGCCs in terms of speaker discriminability. Given feature vectors for N speakers, the J-Ratio can be computed using Eqs. (6) and (7):

$$S_w = \frac{1}{N}\sum_{i=1}^{N} R_i \quad S_b = \frac{1}{N}\sum_{i=1}^{N}(M_i - M_o)(M_i - M_o)^\top \quad (6)$$

$$J = \text{trace}\{(S_b + S_w)^{-1} S_b\}, \quad (7)$$

where S_w is the within-class scatter matrix, S_b is the between-class scatter matrix, M_i is the mean vector for the i^{th} speaker, M_o is the mean of all M_is, and R_i is the covariance matrix for the i^{th} speaker (a higher J-Ratio means better separation). Table 1 shows the J-Ratio for different feature sets; it leads us to two important observations. (1) SGCCs offer better separation than MFCCs. This is partly attributable to the stationarity of subglottal acoustics and the low within-class scatter that results from it. Despite the moderate correlations achieved by the MMSE estimator (Figure 3(a)), estimated SGCCs are comparable in performance to actual SGCCs. This suggests again that the discriminatory power of estimated SGCCs is partly due to the speaker-dependent nature of the estimator (Figure 3(b)). (2) SGCCs are complementary to MFCCs, as reflected by the significantly higher J-Ratios for the combined feature sets. Note that SGCCs are simply concatenated with MFCCs for this analysis; for speaker recognition experiments, we follow the score-combination framework described in Section 2.

4. Speaker recognition experiments

The acoustic models for SID and SV are simple GMMs (as in [1]) and UBM-adapted GMMs (as in [2]), respectively. Given enrollment data, speech segments are first detected using the algorithm proposed in [21]. MFCCs x_0–x_{25} are extracted from the detected speech segments using a 20 ms Hamming window, a 10 ms frame shift, and a 26-channel Mel filter bank. Non-vowel speech frames must be discarded for SGCC estimation since the MMSE estimator is trained on isolated vowels only. Instead of using a vowel detector, we simply retain all speech frames that are strongly voiced. A normalized autocorrelation peak value of 0.6 is used as the threshold to detect strongly-voiced frames. Using MFCCs x_1–x_{25}, SGCCs y_1–y_{25} are estimated and used to train the GMMs of the SGCC system. To train the MFCC GMMs, we use x_0–x_{12} and their first- and second-order derivatives. Note that feature concatenation is not possible here—MFCCs are computed for all speech frames whereas SGCCs are computed for strongly-voiced frames, only.

Given a test utterance, MFCCs and SGCCs are computed as described above. The features are scored with their respective models to obtain two sets of scores (see Figure 2). The scores are log likelihoods for SID and log likelihood ratios for SV. Each set of scores is normalized to the range [0,1]; this is essential before score combination since MFCC and SGCC scores are generally observed to have different dynamic ranges. The scores from the two systems are combined in a weighted fashion such that the weights (non-negative) sum to 1. The combined scores are used to make a decision. Note that the score-combination procedure is not rigorously optimized using separate development and evaluation sets; our focus is more on answering the question as to whether or not SGCCs are beneficial to SID and SV. The results reported in this paper could therefore be a little more optimistic than what we would observe in practice.

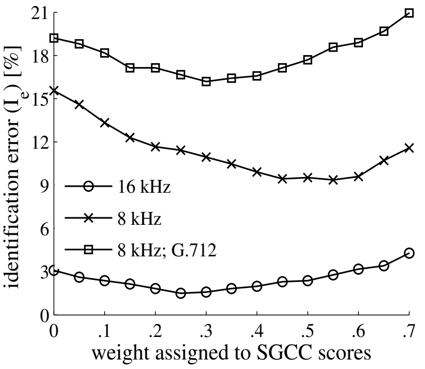

Figure 4: Percent identification error (I_e) as a function of SGCC weight (0 weight = MFCCs only) for the TIMIT database.

Data	Baseline system	Best combined system	Best SGCC weight
16 kHz	3.09	1.51 (51.1%)	0.25
8 kHz	15.56	9.37 (37.8%)	0.55
8 kHz; G.712	19.21	16.19 (15.7%)	0.30

Table 2: Percent identification errors for the TIMIT database in three different conditions, for the baseline (MFCC-only) and the best combined systems (relative reductions in paranthesis).

4.1. Speaker identification: TIMIT database

TIMIT consists of data (sampled at 16 kHz) from 630 speakers. Each speaker has 10 utterances: 2 “shibboleth” *sa* sentences, 5 phonetically-compact *sx* sentences, and 3 phonetically-diverse *si* sentences [22]. The average utterance length is around 3 seconds. The *sa* sentences are used individually as test trials and the remaining 8 sentences are used for acoustic modeling (as in [1]). MFCCs are modeled with 32-component GMMs and SGCCs are modeled with 16-component GMMs.

SID performance is evaluated in three different conditions: (1) wideband (16 kHz sampling rate), (2) narrowband (8 kHz sampling rate), and (3) filtered narrowband (8 kHz sampling rate; data are band-pass filtered using the ITU-T G.712 characteristic [23], which has a flat frequency response from 300 to 3400 Hz). Note that for the filtered narrowband condition, the MMSE estimator is retrained after applying the G.712 characteristic to the vowel segments in the WashU-UCLA corpus.

Figure 4 shows the percent identification error (I_e) as a function of the weight assigned to SGCCs, for the three evaluation conditions described above. Table 2 summarizes the results for the best combined systems along with the I_e reductions relative to their respective baselines. SGCCs are clearly effective and complementary (the optimal SGCC weight is less than 0.5, on average) to MFCCs, and one of the reasons for this is the short duration of the test utterances.

4.2. Speaker verification: NIST 2008 database

NIST 2008 data (used widely for evaluating SV algorithms) are similar to the filtered narrowband speech of TIMIT, but with significantly higher speaker and channel variability [24]. Segments from the “10-sec” condition (which has 10 second utterances from 1336 speakers) are used for this experiment. Data from 892 speakers (having just one utterance each) are used for UBM training. Data from the remaining 444 speakers (having at least two utterances each) are used for enrollment (one utterance) and evaluation (one utterance). The test trials are set up such that each test segment is claimed to belong to each of the 444 speakers, with only one of them being the target speaker. Hence, there are 197136 trials in total. A 128-component UBM is trained for both MFCC and SGCC systems, and speaker models are obtained via maximum *a posteriori* (MAP) adaptation of the UBMs. A relevance factor of 10 is chosen to adapt the means, covariances and component weights. The MSR Identity Toolbox is used for all experiments [25].

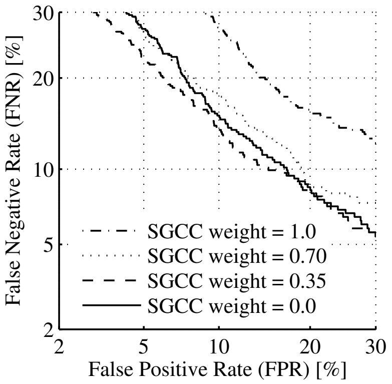

Figure 5: Detection error tradeoff (DET) curves corresponding to different SGCC weights (0 weight = MFCCs only) for the 5 second test trials in the NIST 2008 database.

Note that the above experimental setup is not a standard one. Typically, UBMs are trained on other corpora (Switchboard, Fisher, NIST 2006, etc.), and NIST 2008 data are used for enrollment and evaluation [5, 6, 7]. Nevertheless, our framework serves as a proof-of-concept to demonstrate the efficacy of SGCCs in the presence of speaker and channel variability.

Equal error rate (EER) is used as the performance metric. Evaluation on the 10 second test utterances results in a 4.3% EER reduction for the best combined system (SGCC weight = 0.35), relative to the MFCC-only baseline of 10.59%. The effect of SGCCs is stronger when the test utterances are truncated to 5 seconds each: the best combined system (SGCC weight = 0.35) shows a 10.5% reduction relative to the baseline EER of 12.84%. Detection error tradeoff (DET) curves for the 5 second test trials are shown in Figure 5.

4.3. Discussion

In both SID and SV tasks, the proposed system performs worse than the baseline as the SGCC weight tends to 1. However, the J-Ratio analysis of Sec. 3.2 shows that estimated SGCCs by themselves can provide better speaker separation than MFCCs. This discrepancy could be due to (1) acoustic mismatch between the WashU-UCLA corpus and the speaker recognition corpora, or (2) our simplistic approach to selecting vowel-like frames for SGCC estimation. The above hypotheses could possibly be verified via experiments on a large, phonetically-balanced database (like TIMIT) of speech and subglottal acoustics.

5. Conclusion

We have shown in this paper that SGCCs, estimated in regions of voiced speech activity via a Bayesian MMSE approach, can provide improved speaker-recognition performance when combined with conventional MFCC features at the score level. SID (TIMIT data) and SV experiments (NIST 2008 data) demonstrate the efficacy of SGCCs in different bandwidth conditions and in the presence of speaker and channel variability. The effect of SGCCs is stronger for shorter test utterances.

The results reported here are based on GMMs and UBM-adapted GMMs. Further experiments are required to verify the utility of SGCCs in state-of-the-art *i*-vector systems.

6. References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[3] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[4] C. H. You, K. A. Lee, and H. Li, "An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 49–52, 2009.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[6] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[8] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[9] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Proceedings of ICASSP*, vol. 1, 2006, pp. 665–668.

[10] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3, pp. 455–472, 2005.

[11] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proceedings of Interspeech*, 2001, pp. 2521–2524.

[12] M. Li, J. Kim, P. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on fusion of acoustic and articulatory information," in *Proceedings of Interspeech*, 2013, pp. 1614–1618.

[13] H. Arsikere, G. K. F. Leung, S. M. Lulich, and A. Alwan, "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation," *Speech Communication*, vol. 55, pp. 51–70, 2013.

[14] S. Wang, S. M. Lulich, and A. Alwan, "Automatic detection of the second subglottal resonance and its application to speaker normalization," *Journal of the Acoustical Society of America*, vol. 126, pp. 3268–3277, 2009.

[15] H. Arsikere, S. M. Lulich, and A. Alwan, "Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency," in *Proceedings of ICASSP*, 2013, pp. 7922–7926.

[16] I. Sanchez and H. Pasterkamp, "Tracheal sound spectra depend on body height," *American Review of Respiratory Disease*, vol. 148, pp. 1083–1083, 1993.

[17] S. M. Lulich, J. R. Morton, M. S. Sommers, H. Arsikere, Y.-H. Lee, and A. Alwan, "A new speech corpus for studying subglottal acoustics in speech production, perception, and technology (A)," *Journal of the Acoustical Society of America*, vol. 128, p. 2288, 2010.

[18] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[19] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Proceedings of ICASSP*, 2011, pp. 4624–4627.

[20] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1990.

[21] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[22] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology*, vol. 15, pp. 29–50, 1988.

[23] ITU-T recommendation G.712, "Transmission performance characteristics of pulse code modulation channels," 2001.

[24] A. F. Martin and C. S. Greenberg, "NIST 2008 speaker recognition evaluation: performance across telephone and room microphone channels," in *Proceedings of Interspeech*, 2009, pp. 2579–2582.

[25] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, November 2013.