Robust speaker identification via fusion of subglottal resonances and cepstral features

Jinxi Guo,[a] **Ruochen Yang, Harish Arsikere, and Abeer Alwan**
Department of Electrical Engineering, University of California, Los Angeles, California 90095, USA
lennyguo@g.ucla.edu, yangruoc@gmail.com, hari.arsikere@gmail.com, alwan@ee.ucla.edu

Abstract: This letter investigates the use of subglottal resonances (SGRs) for noise-robust speaker identification (SID). It is motivated by the speaker specificity and stationarity of subglottal acoustics, and the development of noise-robust SGR estimation algorithms which are reliable at low signal-to-noise ratios for large datasets. A two-stage framework is proposed which combines the SGRs with different cepstral features. The cepstral features are used in the first stage to reduce the number of target speakers for a test utterance, and then SGRs are used as complementary second-stage features to conduct identification. Experiments with the TIMIT and NIST 2008 databases show that SGRs, when used in conjunction with power-normalized cepstral coefficients and linear prediction cepstral coefficients, can improve the performance significantly (2%–6% absolute accuracy improvement) across all noise conditions in mismatched situations.

[DDO]
Date Received: November 2, 2016 **Date Accepted:** March 23, 2017

1. Introduction

Robustness of automatic speaker identification (SID) is important for real-world situations. Research has shown that SID systems achieve high accuracy in clean matched conditions, but the performance decreases dramatically for noisy and mismatched conditions (clean training) (Reynolds, 1994; Zhao *et al.*, 2012; Liu and Hansen, 2012). Mel-frequency cepstral coefficients (MFCCs), which are computed by using a Mel-scaled filter-bank, are commonly used features for clean speech SID. However MFCCs are sensitive to noise and the performance degrades significantly in noisy conditions.

In this letter, we investigate the utility of noise-robust subglottal features (capturing the acoustics of the trachea-bronchial airways) for noise robust SID. Since MFCCs are not reliable for SID under noisy conditions, two noise-robust features: high order linear prediction cepstral coefficients (LPCCs) (Reynolds, 1994) and power-normalized cepstral coefficients (PNCCs) (Kim and Stern, 2012) are used as the SID baseline. High order LPCCs represent the smoothed spectrum which is more robust to noise and also keeps speaker specific information; PNCCs use a power-law nonlinearity to suppress small signals, a noise-suppression algorithm based on asymmetric filtering that suppresses background excitation, and a module that models temporal masking. The utility of the noise robust subglottal features [subglottal resonances (SGRs)] in conjunction with PNCCs and LPCCs for SID are studied for two models: Gaussian mixture models adapted from universal background models (UBM-GMM) (Reynolds and Rose, 1995) and i-vector/probabilistic linear discriminate analysis (PLDA) framework (Dehak *et al.*, 2011; Prince, 2007).

We studied the characteristics of SGRs by manually analyzing accelerometer recordings of subglottal acoustics. An automatic algorithm was developed to estimate SGRs from speech signals based on the property that SGRs form the boundary for the front/back and high/low vowel space relative to their relationship with the formant frequencies (Lulich, 2010).

The reasons for the interest in using SGRs as noise robust SID features are as follows. First, the subglottal acoustics are speaker specific owing to some extent to their dependence on body height. We found speech-based SGR estimates to be effective for speaker height estimation (Arsikere *et al.*, 2013a) and adaptation (Arsikere *et al.*, 2013b; Guo *et al.*, 2015). Second, the spectral characteristics of subglottal acoustics are much less variable than the spectral characteristics of the corresponding speech

[a]Author to whom correspondence should be addressed.

signal for a given speaker. The stationary nature can be beneficial especially for limited data and short utterances, which can alleviate the mismatch between training and testing utterances (Guo *et al.*, 2016). Third, previous research has shown that the estimation algorithm of the SGRs is reliable down to 0 dB signal-to-noise ratio (SNR) (Arsikere *et al.*, 2013a). Finally, while the majority of front-end features are related to the supraglottal acoustics, the subglottal features can complement the supraglottal features for SID.

In this letter, we propose a two-stage framework, using LPCCs or PNCCs to reduce the number of target speakers to top N for a given test utterance first and then using SGRs as the complementary features to conduct identification within these N speakers. We evaluate our approach on TIMIT and NIST 2008 databases and show that SGRs offer great complementary information to the baseline systems.

2. Proposed framework

We propose a two-stage framework to fuse the information provided by SGRs and cepstral features.

During the first stage of the proposed system, we use the cepstral features (PNCC and LPCC) as the front-end feature to find the top N most likely speaker models for a test utterance. Within these N speakers, the SGRs are used as new features in the second stage. A multilayer perception (MLP) model (feedforward neural network) is used as the classifier to generate new scores for these N speakers with respect to the corresponding test utterance. The cepstral and SGR scores of the N speakers are then combined in a weighted fashion and the combined scores are used to make the final decision. An overview of the proposed framework is presented in Fig. 1 and the implementation details are provided in Sec. 4.

This two-stage method is adopted here for several reasons. First, since SGRs have negative correlation with speaker height, speakers similar in height might have similar SGRs. Using SGRs to perform identification tasks among a large number of speakers may not be discriminative enough. Pilot experiments showed that, when combining two systems trained using PNCC/LPCC and SGRs individually for all speakers, performance improvement is not significant. However, reducing the number of target speakers to a small number, N, may help SGRs to better discriminate speakers and give useful complimentary information. Furthermore, compared with the traditional single stage combination of two individual systems, the proposed two-stage framework can reduce the latency by limiting the number of evaluation targets in the second stage.

3. SGR estimation

3.1 Estimation algorithm

The SGR estimation algorithm follows the algorithm proposed in Arsikere *et al.* (2013a). The algorithm is based on the following idea: $Sg1$ acts as a boundary between high and low vowels so that two acoustic features characterizing vowel frontness—the Bark difference between $F3$ and $F1$ (denoted as B_{31}) and the Bark difference between $F1$ and $Sg1$ (denoted as $B_{1,s1}$)—are correlated. Similarly, for $Sg2$ estimation, the Bark difference between $F3$ and $F2$ (denoted as B_{32}) was found to be related to the bark difference between $F2$ and $Sg2$ (denoted as $B_{2,s2}$), since both measures characterize vowel backness. Two empirical equations were derived to predict $B_{1,s1}$ and $B_{2,s2}$ using regression models as in Eqs. (1) and (2). $Sg3$ is estimated based on its correlation with $Sg2$ using a first-order linear regression, as in Eq. (3):

$$B_{1,s1} = 0.011(B_{31})^3 - 0.269(B_{31})^2 + 1.322(B_{31}) + 2.455. \quad (1)$$

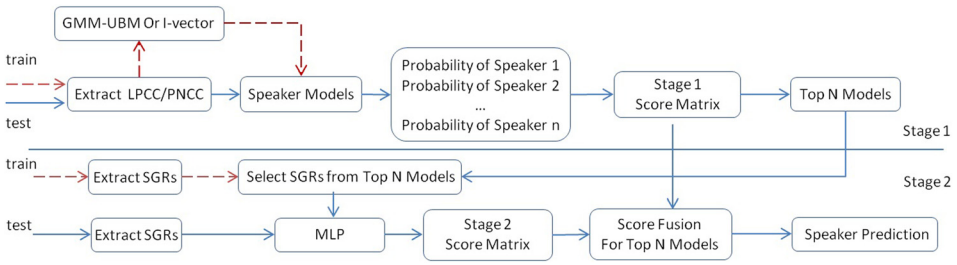

Fig. 1. (Color online) System flow chart.

$$B_{2,s2} = -0.004(B_{32})^3 + 0.134(B_{32})^2 - 1.958(B_{32}) + 6.182, \quad (2)$$

$$Sg3 = 1.079 * Sg2 + 763.676. \quad (3)$$

Given a speech utterance, the first three formants of the voiced frames are estimated using SNACK and voiced activity detection (VAD). The three SGRs can then be estimated using the above empirical relations. Note that formant values used in the equations are in the Bark scale. Since $Sg1$ and $Sg2$ act as boundaries in the vowel space, estimation results are expected to be largely phoneme independent (Arsikere *et al.*, 2013a). For noise-robust SGR estimation, what we need is to make the estimation of noisy-speech formants reasonably close to the corresponding clean-speech formants, even though the absolute values of the estimated formants are not very accurate. In general, SNACK's estimates of formants between clean and noisy speech were found to be close in voiced regions even for low local SNRs (Sjolander, 2004). Therefore, the proposed SGR estimation algorithm is expected to be noise robust.

3.2 Results

Three additive noises (i.e., babble, factory, and pink) collected from the NOISEX-92 database were used for representing different noise conditions. The speech segment was degraded by adding a specific type of noise at SNRs of 5, 10, 15, 20 dB, respectively, using FaNT (Hirsch, 2005).

SGRs for all 630 Speakers in TIMIT are estimated based on one clean utterance per speaker, as well as the same utterance with additive noise of different SNRs. Two speakers (denoted as Speaker 160 and 600) are selected to show the effectiveness of the estimation algorithm. Figure 2 illustrates the comparison of estimated SGRs across all the frames for a selected clean utterance and 12 different noise/SNR combinations for two speakers (1 female and 1 male). The x axis indicates noise conditions: 1, clean; 2–5, Babble 5/10/15/20 dB; 6–9, Factory1 5/10/15/20 dB; and 10–13, Pink 5/10/15/20 dB. Each symbol represents the mean value of each SGR estimate given a noise condition. The bars correspond to the standard deviation (STD) of each estimate. The means of the estimated SGRs for different conditions are similar, and this implies that the estimated SGRs are fairly constant across all test noise conditions. The narrow STD intervals indicate that, given a noise condition, the estimation algorithm is quite noise robust. Moreover, comparing the two speakers in Fig. 2, it is clear that their estimated SGRs are distinctly different.

To further quantify the SGR estimation accuracy for all speakers, the average root mean square error (denoted as RMSE_{avg}) is used, which measures the differences between the SGRs estimated from the clean and corresponding noisy utterances. Table 1 shows the RMSE_{avg} of the averaged SGR estimates under all noise types compared to clean for a given SNR. As expected, the RMSE of SGRs is fairly small even for low SNRs. Table 2 also demonstrates that the overall RMSE_{avg} across all SNRs is small for a given noise type and the estimation error for the babble noise case is smaller than factory and pink noise. The results confirm the robustness of the SGR estimation algorithm on a large SID data set. Similarly, the experiment on NIST SRE08 yields similar results, which shows relatively small estimation errors between clean and corresponding noisy conditions. Thus, it is beneficial to incorporate the SGRs for noise robust SID tasks.

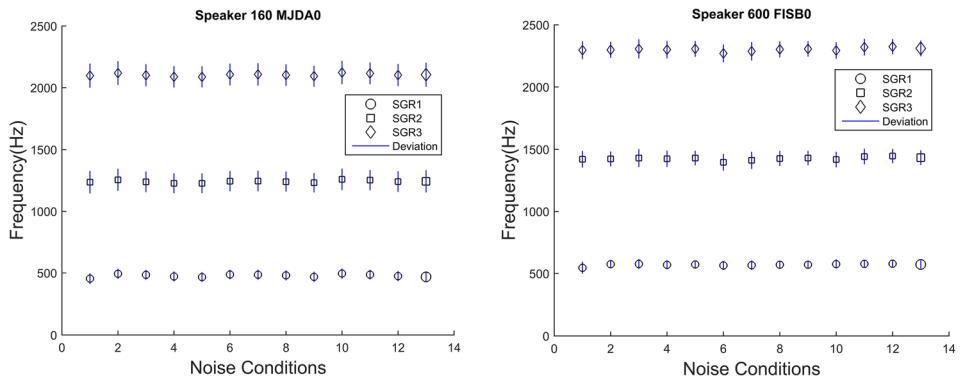

Fig. 2. (Color online) SGR estimation for speaker 160 (left) and speaker 600 (right) (the x axis indicates noise conditions: 1, clean; 2–5, Babble 5/10/15/20 dB; 6–9, Factory1 5/10/15/20 dB; and 10–13, Pink 5/10/15/20 dB).

Table 1. Overall RMSE (in Hz) of SGRs under several SNRs (TIMIT).

	5 dB	10 dB	15 dB	20 dB
SGR1	41.0753	30.0358	20.3100	11.6578
SGR2	77.8863	58.1728	40.9838	25.5018
SGR3	83.7503	62.4795	42.8535	27.2275

4. SID experiments and results

All experiments were conducted under mismatched conditions with clean training utterances evaluated against noisy test utterances.

The TIMIT SID acoustic models are UBM GMMs (Reynolds and Rose, 1995). Since TIMIT only has files with very short utterances, the UBM GMM framework is used here. On the other hand, the state of the art i-vector/PLDA model is used on the NIST08 dataset. Given the enrollment data, speech segments are first detected using a statistic-based VAD algorithm (Sohn *et al.*, 1999) to discard non-speech frames. In total four front-end features are extracted: LPCCs, PNCCs, SGRs, and MFCCs. For MFCCs and PNCCs, we use the first 20 coefficients and their first- and second-order derivatives, resulting in 60-dimensional features. For LPCCs, the first 24 coefficients are used for our experiment (adding derivatives did not result in improvement). Note that all the cepstral features are computed for all speech frames whereas SGRs are computed for voiced frames only.

Given a test utterance, the cepstral features and SGRs are computed as described above. The cepstral features are scored with their respective models. The scores are log likelihoods and are normalized using $S_{i_{\text{norm}}} = S_i - S_{\min}/S_{\max} - S_{\min}$. The top N highest scoring speakers are selected for the test utterance. The selected top N scores are further normalized to a score between 0 and 1 (by dividing by the sum of the scores). For these N speakers, SGRs are used as the new features and MLP as the new classifier. One hidden layer with sigmoid activation function is used for training, and the softmax function is adopted to get the normalized scores from the output layer. The scores from the two stages are combined in a weighted fashion, and the weights are two positive scalars and summed up to one. Weights are determined empirically across the whole test set such that the highest accuracy is achieved. The combined scores are used to make a decision. Note that, since the training of the shallow neural network model (MLP) for the selected three speakers using three SGRs is very time-efficient, for online testing in real applications, the MLP can be trained online after the top N candidates are selected from the pre-trained first-stage model.

4.1 SID on TIMIT database

TIMIT consists of 10 utterances spoken by 630 speakers, with a sampling rate 16 kHz (Garofolo, 1988). The average utterance length is around 3 s. One of the ten utterances is used as the test trial for each speaker and the remaining nine sentences are used for acoustic modeling. Cepstral features are modeled with 128-component GMMs. SID performance is evaluated in 12 different SNR conditions. The number of speakers, N, chosen for the second stage is set to 3, 5, and 10. Pilot results indicated that there is no significant advantage using a larger N. Therefore we set N to 3 for fast training and testing; evaluation results are shown in Table 3. Table 3 shows the SID results for the MFCC, PNCC, LPCC baseline systems, and PNCC+SGRs, LPCC+SGRs combined systems with best weights. The percentage of the predication accuracy for SID is used as the metric. Since the MFCC baseline was low in noisy conditions, we did not evaluate it with SGRs. The combined feature systems perform the best across all noise conditions, and give relatively bigger improvement for pink, factory, and low-SNR babble noise. Figure 3 shows the weight ratios of SGRs across SNRs for pink noise, when

Table 2. Overall RMSE (in Hz) of SGRs under several noise types (TIMIT).

	Babble	Factory1	Pink
SGR1	19.3118	25.4771	31.9324
SGR2	37.5271	49.8199	63.0521
SGR3	40.4917	53.7557	68.0332

Table 3. SID accuracies (%) under different noise and SNR combinations for TIMIT (boldface numbers indicate best results).

	MFCC	PNCC	LPCC	PNCC+SGRs	LPCC+SGRs
Babble					
5 dB	46.7	53.5	64.7	56.4	**70.8**
10 dB	85.2	80.6	93	82.4	**94.1**
15 dB	95.7	89.8	98	92.0	**98.4**
20 dB	97.7	91.7	99.2	92.4	**99.3**
Factory					
5 dB	14.3	24	23.1	**30**	27.8
10 dB	41.9	59.3	56.6	**63.4**	60.5
15 dB	73.8	83.6	81.9	84.7	**86.9**
20 dB	92.5	91.4	95.8	92.4	**97.3**
Pink					
5 dB	4.6	18.9	7.4	**22.1**	10.9
10 dB	17	34.7	26.1	**39.9**	31.6
15 dB	42.4	57.9	52.8	**61.3**	59.7
20 dB	71.9	80.6	78.8	83.2	**85.2**

combining with LPCCs. SGR weights increase as SNR decreases, which indicates that SGRs are more effective in low-SNR.

4.2 SID on NIST 2008 database

NIST 2008 data are widely used for evaluating speaker verification (SV) algorithms (Martin and Greenberg, 2009). Compared with TIMIT, it has higher speaker and channel variability. Note that unlike the standard SV task, in this letter we only focus on the SID task and demonstrate the efficacy of the SGRs in the presence of larger speaker and channel variability. Therefore, we randomly chose 947 speakers from the evaluation dataset (3conv part of the training set). Each speaker participates in three telephone conversations. For each utterance in the telephone conversations (approximately 5 min long), a 10 s segment is extracted as the test utterance and the remaining part is used for training the speaker model.

Since our experiment is only concerned with the closed-set SID task, the training data can be used to set up the i-vector/PLDA system. A gender independent UBM of 1024 GMM components was built. A total variability matrix T of 400 factors was used and the dimension of the resultant i-vector was further reduced via PLDA modeling with 200 latent components.

Table 4 summarizes the results for the performance of the proposed combined features system and the baselines. The baseline for 20 dB SNR and matched clean conditions are above 98% and the improvement is small; we show results for 5/10/15 dB here. Similar to the TIMIT experiment, the combined features system outperforms the other baselines in all noise conditions, which shows the significant complementary

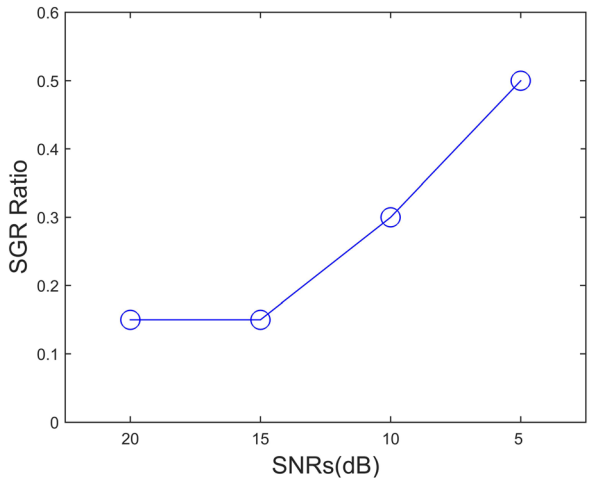

Fig. 3. (Color online) SGR ratio across SNRs for pink noise.

Table 4. SID accuracies (%) under different noise and SNR combinations for NIST SRE 08 (boldface numbers indicate best results).

	MFCC	PNCC	LPCC	PNCC+SGRs	LPCC+SGRs
Babble					
5 dB	16.6	46.2	37.9	**50.6**	43.6
10 dB	45.7	76.1	70.3	**80.5**	76.5
15 dB	75.0	89.8	90.5	92.0	**92.8**
Factory					
5 dB	20.5	44.5	40.5	**49.6**	46.1
10 dB	54.9	75.2	74.3	**79.4**	78.2
15 dB	84.3	89.6	93.9	91.4	**95.2**
Pink					
5 dB	17.6	47.7	24.0	**53.2**	30
10 dB	53.7	77.8	63.3	**81.4**	68.2
15 dB	85.2	90.7	89.5	**92.8**	91.1

effect for SGRs to the baseline cepstral features. The best SGR weights for the combined systems also increase for a low-SNR condition.

To further analyze how SGRs help improve the SID accuracy, we check the top three cepstral scores for the clean test utterances and the corresponding noisy utterances. As expected, the scores for noisy data are not prominent for a certain speaker, since cepstral features tend to be suboptimal in the presence of noise, which leads to greater confusion among acoustically-similar speakers. Since the SGRs are more noise robust and speaker specific, when we fuse the cepstral score with scores from SGRs, the fused scores become more prominent for the target speaker, which indicates that the SGRs actually help the decision making.

Another reason why using SGRs give improvement to the SID system is that, since SGRs are more stationary compared with standard cepstral features which represent the vocal tract information, it can help to alleviate the phoneme mismatch problem between train and test utterances when only short utterances are available (e.g., 10 s).

5. Conclusions

In this letter, a two-stage noise robust SID system is proposed to demonstrate the efficacy of the SGRs as complementary noise-robust features. SID experiments on TIMIT and NIST 2008 database demonstrates that SGRs can provide complimentary speaker information to noise robust features, such as PNCCs and LPCCs.

Acknowledgment

This work was supported in part by NSF Grant No. 0905381.

References and links

Arsikere, H., Leung, G. K. F., Lulich, S. M., and Alwan, A. (**2013a**). "Automatic estimation of the first three subglottal resonances from adults speech signals with application to speaker height estimation," Speech Commun. **55**(1), 51–70.

Arsikere, H., Lulich, S. M., and Alwan, A. (**2013b**). "Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 7922–7926.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (**2011**). "Front-end factor analysis for speaker verification," IEEE Trans. Speech Audio Process. **19**(4), 788–798.

Guo, J., Paturi, R., Yeung, G., Lulich, S. M., Arsikere, H., and Alwan, A. (**2015**). "Age-dependent height estimation and speaker normalization for children's speech using the first three subglottal resonances," in *Interspeech*, pp. 1665–1669.

Guo, J., Yeung, G., Muralidharan, D., Arsikere, H., Afshan, A., and Alwan, A. (**2016**). "Speaker verification using short utterances with DNN-based estimation of subglottal acoustic features," in *Interspeech*, pp. 2219–2222.

Garofolo, J. S. (**1988**). "DARPA TIMIT acoustic-phonetic speech database," Natl. Inst. Stand. Technol. **15**, 29–50.

Hirsch, H. G. (**2005**). FaNT-Filtering and Noise Adding Tool.

Kim, C., and Stern, R. M. (**2012**). "Power-normalized Cepstral coefficients (PNCC) for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4101–4104.

Liu, G., and Hansen, H. L. (**2012**). "Robust feature front-end for speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) (March 25–30), pp. 4233–4236.

Lulich, S. M. (**2010**). "Subglottal resonances and distinctive features," J. Phonetics **38**(1), 20–32.

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

Prince, S. J. D. (**2007**). "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8.

Reynolds, D. A. (**1994**). "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Process. **2**(4), 639–643.

Reynolds, D. A., and Rose, R. C. (**1995**). "Robust text independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**(1), 72–83.

Sjolander, K. (**2004**). "The Snack Sound Toolkit," http://www.speech.kth.se/snack/.

Sohn, J., Kim, N. S., and Sung, W. (**1999**). "A statistical model-based voice activity detection," IEEE Signal Process. Lett. **6**(1), 1–3.

Zhao, X., Shao, Y., and Wang, D. (**2012**). "CASA-based robust speaker identification," IEEE Trans. Speech Audio Process. **20**(5), 1608–1616.