SOFTSAD: INTEGRATED FRAME-BASED SPEECH CONFIDENCE FOR SPEAKER RECOGNITION

Mitchell McLaren, Martin Graciarena, Yun Lei

Speech Technology and Research Laboratory, SRI International, California, USA

{mitch,martin,yunlei}@speech.sri.com

ABSTRACT

In this paper we propose softSAD: the direct integration of speech posteriors into a speaker recognition system as an alternative to using speech activity detection (SAD). Motivated by the need to use audio from short recordings more efficiently, softSAD removes the need to discard audio using speech/non-speech decisions based on a threshold as done with SAD. Instead, softSAD explicitly integrates into the Baum-Welch statistics a speech posterior for each frame. We compare softSAD and SAD in mismatched conditions by evaluating a system developed for the National Institute for Standards and Technology (NIST) 2012 speaker recognition evaluation (SRE) on the short test conditions of the channel-degraded Robust Automatic Transcription of Speech (RATS) speaker identification task (and vice versa). We demonstrate that softSAD provides benefit over SAD for short test audio in mismatched conditions.

Index Terms— Speech activity detection, speaker identification, unseen conditions, mismatched conditions.

1. INTRODUCTION

Speech activity detection (SAD) is fundamental to almost all speech processing applications, including speech recognition, language recognition, and the focus in this work, speaker identification (SID). SAD can be viewed as an audio pre-processing module that filters frames (i.e., 25ms windows of overlapping audio) from the audio stream that are not expected to provide information for the end task (i.e., SID). The extent to which non-speech audio is filtered is typically tuned with a threshold; this threshold may differ with application. For instance, in the case of speech recognition where voiceless sounds are informative for understanding, a low threshold might be used. In contrast, SID might derive benefit from a more stringent threshold to obtain a higher relative proportion of voiced sounds that are rich in speaker information.

The most popular methods of implementing SAD involve Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and neural networks [1]. In this work, we concentrate on the GMM-based approach to SAD, which is shown to be highly successful in both NIST SRE'12 [2] and the Defense Advanced Research Projects Agency (DARPA) Robust Automatic Transcription of Speech (RATS) SID task [3, 4]. This approach uses GMMs to model spectral features and obtain speech/non-speech likelihood ratios to which a threshold is applied to obtain a binary detection value. Irrespective of the modeling approach used, a development dataset is required to learn the SAD model. Consequently, the success of the SAD model depends both on the tuned threshold and the ability of the development data to reflect end use conditions.

In this work, we concentrate on improving the robustness of speaker recognition under short duration, mismatched train/test conditions by reducing the dependence of SAD on the tuned threshold. Traditionally, a SID system calculates Baum-Welch statistics by equally weighting each speech frame found using the binary detection of SAD. We propose to remove this detection phase, and instead use every audio frame after weighting it by its speech posterior; or a confidence measure for speech. We term this approach *softSAD*. SoftSAD attempts to utilize all information in the audio stream while placing emphasis on the more speech-like frames. We anticipate that the benefits of this approach (over conventional SAD) include more efficient use of limited testing or system training data, and robustness to evaluation conditions that are greatly mismatched to the SAD training conditions, since information from audio with low speech posteriors will still be used instead of being discarded as non-speech based on a threshold.

2. MODELING SPEECH ACTIVITY

Modeling speech activity for speaker recognition typically involves determining which frames of audio contain speech. This is a binary detection task that reduces the amount of audio frames to be processed by the system and presents to the system audio that is rich in information for the task at hand. Accurate speech selection is crucial for speaker recognition, as shown in [2, 5]. In this section we discuss the pros and cons of the common approach to modeling speech activity and propose speech activity posteriors (softSAD) to improve the robustness of speech activity modeling for short duration detection tasks in unseen conditions.

2.1. Speech Activity Detection (SAD)

SAD, like many detection tasks, involves modeling speech as observed in a development dataset, then deciding which frames of processed audio are speech and should, therefore, be processed by the system. Modeling approaches have focused largely on GMMs, HMMs, neural networks and more recently, deep neural networks [1]. Neural networks and HMMs are ideal when it is useful to retain low-energy voiceless speech for natural language processing or user intelligibility [6]. GMMs, on the other hand, are simple and elegant for SID where understanding of speech content is not essential but the localization of voiced, high-energy frames is more critical [2, 5]. While not considered in this work, of note is the ability

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its contracting agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. "A" (Approved for Public Release, Distribution Unlimited).

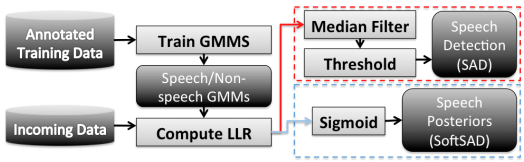

Fig. 1. The GMM-based approach to computing smoothed speech/non-speech likelihood ratios, and the subsequent SAD and softSAD processing stages.

of score-level fusion of SID systems implemented on top of different SAD models to provide some robustness to heavily degraded conditions [4, 1]. The GMM-based approach was commonplace in many submissions to the recent NIST SRE's [2] and DARPA RATS SID task [7, 4]. Based on the SRI team's developments under SRE [2] and the SCENIC team for the RATS SID task [4], we focus on the GMM-based approach to SAD. For GMM-based SAD, a threshold is applied to speech/non-speech likelihood ratios.

2.1.1. Generating Speech/Non-speech Likelihood Ratios

Modeling of speech activity using the GMM-based approach involves training of a speech and non-speech GMM from a development data set of features such as Mel frequency cepstral coefficients (MFCC). The likelihood ratio (LLR) of speech vs. non-speech for incoming audio frames is calculated using the trained models. To obtain speech activity detection (SAD), these LLRs are first smoothed using a median filter spanning 410ms, then thresholded to obtain a binary speech or non-speech flag associated with each audio frame. Figure 1 illustrates the stages involved in the simple GMM-based speech/non-speech modeling and subsequent SAD and softSAD approaches considered in this work. In this study, SAD performance on both corpora considered was obtained using 1024 Gaussian component speech/non-speech models based on 20D MFCCs for SRE'12, or power-normalized cepstral coefficients (PNCC) for RATS, with deltas and double-deltas appended.

2.2. Speech Activity Posteriors

The above SAD process involves production of smoothed likelihood ratios of speech/non-speech from which a detection is made based on a tuned threshold. We propose to directly use a transformation of the LLRs as computed in Section 2.1.1 in the Baum-Welch statistics calculation, thus avoiding the need to make a speech/non-speech decision altogether. We first convert the LLRs to speech posteriors through the application of a sigmoid function,

$$\sigma(llr) = \frac{1}{1 + e^{-\alpha(llr+\beta)}}. \quad (1)$$

The sigmoid parameters α and β are tuned later in the study. The zero- and first-order Baum-Welch statistics ($\mathbf{N}$ and $\mathbf{F}$, respectively) can then be calculated as:

$$\mathbf{N}_k = \sum_t \sigma(llr_t)\gamma_{kt} \quad and \quad \mathbf{F}_k = \sum_t \sigma(llr_t)\gamma_{kt}x_t$$

where x_t represents the feature vector extracted from the audio at time t, and γ_{kt} the alignments for x_t from the k-th component of the universal background model (UBM).

Figure 2 illustrates how different parameters in the sigmoid function affect the transformation of LLRs to posteriors. If α is set toward infinity and β to the SAD threshold, the same output as SAD will be obtained. By using a more tapered change in posteriors, we

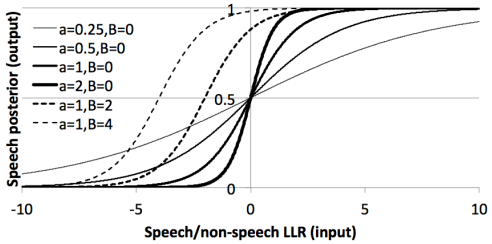

Fig. 2. The effect of varying sigmoid alpha and beta parameters on the LLR to speech posterior transformation.

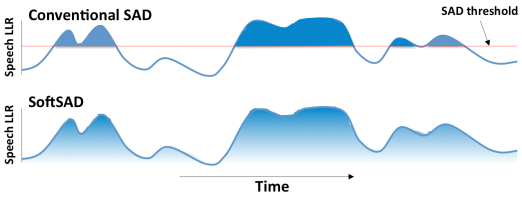

Fig. 3. SAD applies a threshold to determine speech frames which are equally weighted with respect to the SID system. The proposed softSAD method avoids a threshold and uses all speech frames weighted by their corresponding speech posteriors.

can weight each frame according to how well it represents speech according to the speech/non-speech models. Figure 3 provides a pictorial comparison between SAD and softSAD approaches in which softSAD weights all frames according to their speech LLRs.

A number of benefits are anticipated using softSAD over SAD. Firstly, softSAD attempts to utilize all speech information in the audio stream, which should in turn improve low-resource system training or low-resource and short audio enrollment/testing conditions. Second, the ability to place more emphasis on the most speech-rich audio, instead of treating all speech frames equally, may enable the system to more readily exploit the speaker information in high-energy voiced audio. Finally, the combination of the weighting process and the use of all audio frames is expected to provide improved robustness to speech activity modeling, and likewise to speaker recognition performance in severely mismatched conditions, since rather than removing frames with low speech LLRs as perceived by tuned SAD models, the soft posteriors will retain this information in the system. The cost of softSAD over SAD is the potential for unnecessary computation of largely non-speech regions of audio. While not investigated here, it would be intuitive to threshold speech posteriors at a very low value to reduce computation.

3. SEVERELY MISMATCHED DATA SOURCES

Two sources of severely mismatched data are used in this study; the NIST SRE'12 and RATS SID data. Table 1 details the major factors that differentiate these two datasets. In addition, the majority of microphone audio from the SRE'12 set includes both the speaker of interest and an interlocutor. We have previously shown the need to remove the cross-talk from these channels [2]. For the purpose of this study, in which we desire an analysis free of the variability associated with cross-talk detection, we fix the interlocutor speech as detected with the tuned SAD system and discount these audio frames from system analysis for all experiments.

The independent development of systems targeted toward two severely mismatched datasets can result in significant differences in system design, as highlighted in Section 4. Major differences in this work include the features (MFCC vs PNCC), post-processing of fea-

Table 1. Characteristics of the severely mismatched NIST SRE'12 and DARPA RATS SID corpora considered in this work.

SRE'12
- *Channels*: clean and re-noised microphone/telephone
- *Noise*: additive HVAC/babble, environment noise
- *Duration (system train/eval)*: 5-8 mins / 30-200 seconds
- *Gender distribution*: 57% female, 43% male
- *Evaluation language*: English

RATS
- *Channels*: clean + 8 heavily degraded transmission channels
- *Noise*: telephone (clean) and push-to-talk channels
- *Duration (system train/eval)*: 10-15 mins / 10 seconds
- *Gender distribution* : 31% female, 69% male
- *Evaluation languages*: Lev. Arabic, Dari, Farsi, Pashto, Urdu

tures (appended deltas and double deltas vs. rank-DCT coefficients), and dependence vs. independence on gender and channel-awareness associated with the SRE'12 and RATS training data. It should be noted that results presented on the RATS corpus exclude cross-gender trials so as to facilitate the evaluation of the mismatched gender-dependent SRE'12 system. This differs from the official RATS protocol that includes such trials, and consequently, the results reported in this study are significantly worse than those presented in our previous work [4].

4. PROTOCOL AND SYSTEM CONFIGURATION

4.1. Tuning Protocol

Experiments were conducted in the following manner. First, both SRE'12 and RATS systems and their corresponding SAD components were tuned independently. The system models were then fixed before introduction of softSAD, which was tuned to optimize SID performance by using softSAD on the enrollment and test data of each corpus. The system was then re-trained using the softSAD throughout to produce the optimized SRE and RATS systems.

4.2. System Configurations

All systems are based on the i-vector/probabilistic linear discriminant analysis (PLDA) framework [8, 9]. UBMs consisted of 2048 components, i-vectors of 600-dimensions and i-vectors were length-normalized and LDA-reduced prior to PLDA.

All features used in the SID components were mean- and variance-normalized (MVN) across speech frames detected via SAD. Specifically in the case of softSAD, SID performance was found to be more stable on the development set when MVN was applied in the same manner as SAD instead of normalizing by a *weighted* mean and variance statistics that would seem a better fit to softSAD. This process has the drawback of requiring at least some speech to be detected by SAD in order to be processed, with limited SAD output reducing feature stability.

SRE'12 System: MFCCs with deltas and double deltas were used for both SID and speech activity modeling. For speech activity modeling, c0 normalization was employed by subtracting the maximum of the first cepstral coefficient (c0) from c0 of the features from a given audio file. This method was found to be particularly beneficial for microphone audio. Gender-dependent systems were trained in the same manner as our SRE'12 submission [2]. A subset of 8,000 clean speech samples were used to train a 2048-component

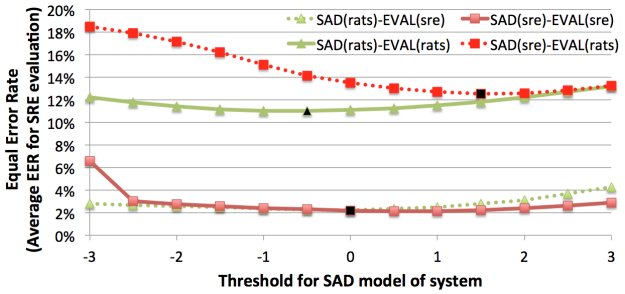

Fig. 4. The effect of varying the SAD threshold on SID performance when evaluated on conditions matched (solid lines) and mismatched (dashed lines) to the SAD model. Both RATS and SRE evaluations are plotted as EVAL(rats) and EVAL(sre), respectively.

UBM for each gender. The 600D i-vector subspace was trained using 51,224 samples; the 350D LDA reduction matrix and full-rank PLDA were trained using using an extended dataset of 62,277 samples (26k of which were re-noised). Evaluation was performed on pooled male and female trials of the five *extended* conditions defined by NIST based with performance reported in terms of equal error rate (EER) and Cprimary [10], the latter being an average of two operating points.

RATS System: PNCCs [11] were used for noise robustness in both speech activity modeling and SID components. While deltas and double deltas were applied for speech modeling, the PNCCs were converted to 100-dimensional rankDCT features (see our companion paper on DCT coefficients for speaker recognition [12]) for SID modeling, which extended our previous work on DCT coefficients in [13]. The data-driven rankDCT features used a subset of 1000 randomly selected training segments which were evenly distributed across channels to learn the 100 coefficients for selection as features from the 2D-DCT matrix obtained by applying a moving DCT window over PNCC features. These DCT-based features were found to provide the best performance for our 2014 submission to the DARPA RATS SID task. A gender-independent system was trained similarly to [14] using 55,982 transmissions including clean source recordings for the UBM and i-vector subspace. This dataset had a 10, 30 and 120 second segment extracted from each transmission and source audio for the training of the LDA and PLDA models.

5. RESULTS

We commence by illustrating the effect of tuning the SAD threshold on development data and the need for improved SAD generalization. SoftSAD is then introduced and independently tuned on the enroll and test sets of the corpora before retraining each individual softSAD-based SID system. Both SAD and softSAD-based SID systems are finally evaluated on the mismatched corpus with short durations to observe the ability of each to generalize to unseen data.

5.1. The SAD Generalization Issue

Both SRE and RATS system models were tuned during the NIST SRE'12 and RATS SID Phase III development phases. We commence with the SAD models resulting from this development effort. In this section, we aim to observe the effect of changing the SAD threshold on the development set as well as the alternate corpus.

Figure 4 illustrates the difference in performance when varying the SAD speech/non-speech detection threshold on both corpora us-

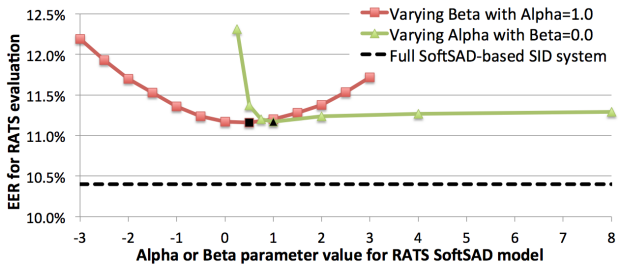

Fig. 5. The effect of varying the RATS softSAD parameters on RATS performance (matched condition) for a SAD-based system. The β is analogous to the SAD threshold when α approaches infinity. The black dashed line indicates the performance after retraining the whole SID system to be aware of the tuned softSAD parameters.

ing SAD models matched to the conditions (solid line) and trained on the alternate corpora (dashed line). The SID models were matched to the evaluation data so as to allow accurate assessment of the impact of SAD. Figure 4 shows that the average EER performance obtained on SRE'12 when using matched SAD performs similar to that of mismatched SAD when operating around a threshold of 0.0. Contrasting trends where found for the RATS data evaluation. The performance degradation due to switching SAD model in this case was more evident, with the best thresholds varying between -0.5 and 1.5.

These results suggest that tuned thresholds for SAD models exposed to a wide variety of conditions generalize well to a subset of those conditions (i.e., RATS model to cleaner SRE'12 audio) when considering the EER operating point. Secondly, the use of SAD models and threshold tuned on a restricted set of conditions does not generalize particularly well (SRE SAD in the RATS system). SoftSAD attempts to address this issue.

5.2. SoftSAD Tuning

In this section, we tune the softSAD parameters on the matched development datasets. This is done by fixing the system models (trained using the best SAD threshold form Section 5.1) and evaluating the development audio using a variety of softSAD parameters. Figure 5 shows how performance varies with these parameters on the RATS dataset. A $\alpha = 1.0$ and $\beta = -0.5$ was optimal for RATS while $\beta = -1.0$ was better for SRE (not shown). From the figure, we can conclude that varying α, the softness factor, between 0.75 at 8 has limited effect on performance. We see, however, that β does affect performance. Although analogous to the SAD threshold, the varying of β within ± 2 from the optimal showed less variation in performance for SoftSAD than for SAD. It is anticipated that this characteristic will allow SoftSAD to generalize more readily than SAD. The same trends were observed in the corresponding SRE plot (not shown due to space limitations) with only a subtle shift in β.

Given the tuned softSAD configuration, both SRE and RATS SID system models were retrained to incorporate softSAD instead of SAD. This approach has a two-fold benefit: it includes additional system training data (i.e., more frames to analyze) and allows the system to better exploit the more speech-like frames. Performance after re-training is also depicted in Figure 5 as a black dashed line, with a significant 13% improvement in system performance on RATS data, while only marginal gains of 3% were found for the SRE corpus (not shown). These results demonstrate that even in matched conditions, softSAD provides benefit over SAD at the cost of additional frames processing.

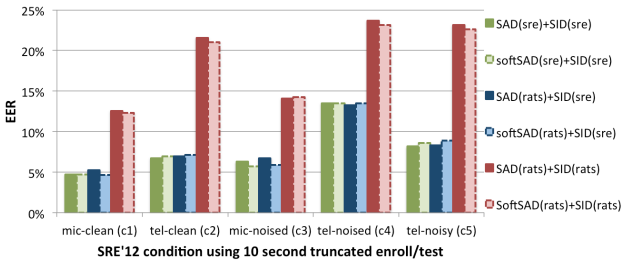

Fig. 6. Evaluation of SRE'12 extended protocol conditions using matched (SRE-SAD/softSAD) and mismatched (RATS-SAD/softSAD) systems with both SAD and softSAD approaches.

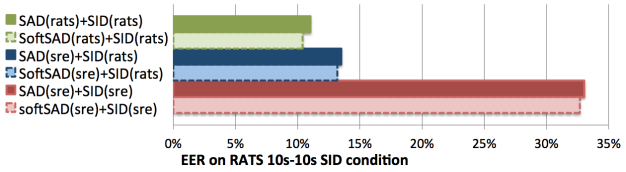

Fig. 7. Evaluation of RATS 10s-10s SID task using matched (RATS-SAD/softSAD) and mismatched (SRE-SAD/softSAD) systems.

5.3. Speech Activity Generalization to Unseen Conditions

This section aims to determine whether the tuned speech activity detection/posterior approaches and corresponding systems generalize well to severely mismatched and short duration data. Specifically, both SAD and softSAD SRE SID systems, tuned using SRE'12 development data, are evaluated on the RATS SID task (and vice versa). Figure 6 and Figure 7 detail results from these experiments. For a more direct comparison to the 10s-10s RATS SID task, all SRE'12 enroll and test segments were truncated to 10 seconds of speech according to SAD outputs. For each evaluation corpus, three conditions are detailed each with SAD or softSAD: SAD and SID models matched (green) or mismatched (red) to evaluation conditions, or only the SAD component mismatched to evaluation conditions (blue). Results in Figure 7 show that softSAD consistently outperforms SAD in both matched and mismatched conditions in the short, degraded conditions of RATS. When using RATS systems for the SRE evaluation in Figure 6 (red bars), performance was marginally better from softSAD, but otherwise comparable. While not shown here, we also evaluated SRE'12 using full length audio segments which resulted in comparable results between SAD and softSAD across the board. Based on these results, the application of softSAD in place of SAD is most beneficial in mismatched, low resource conditions.

6. CONCLUSIONS

We proposed the use of speech activity posteriors (softSAD) to replace traditional speech activity detection (SAD) for the purpose of speaker recognition. SoftSAD integrates the frame speech posterior into the Baum-Welch statistics, thereby utilizing all frames of the audio with different contributions to the final statistics. Speech/non-speech likelihood ratios were converted to posteriors using a sigmoid function. Through a series of experiments on both SRE'12 and RATS SID data, we showed that a tuned SAD threshold does not generalize particularly well to severely mismatched conditions, and in both matched and mismatched conditions, the proposed softSAD was a more robust alternative under degraded, low resource conditions. Future work will consider alternate methods of producing speech posteriors and alternate confidence measures, such as localized audio quality, that may benefit detection tasks.

7. REFERENCES

[1] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program," in *Proc. Interspeech*, 2013, pp. 3497–3501.

[2] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," in *Proc. Interpseech*, 2013.

[3] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. Odyssey-The Speaker and Language Recognition Workshop*, 2012.

[4] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, and Y. Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion," in *Proc. ICASSP*, 2013, pp. 6773–6777.

[5] M. McLaren and D. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *Proc. NIST SRE 2011 workshop, Atlanta, US*, 2011.

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

[7] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program.," in *Proc. Interspeech*, 2012.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Speech and Audio Processing*, vol. 19, pp. 788–798, 2011.

[9] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*. IEEE, 2007, pp. 1–8.

[10] *The NIST Year 2012 Speaker Recognition Evaluation Plan*, 2012, `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf`.

[11] C. Kim and R.M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP*. IEEE, 2012, pp. 4101–4104.

[12] M. McLaren and Y. Lei, "Improved speaker recognition using DCT coefficients as features," in *Proc. ICASSP (submitted)*, 2015.

[13] M. McLaren, N. Scheffer, L. Ferrer, and Y. Lei, "Effective use of DCTs for contextualizing features for speaker recognition," in *Proc. ICASSP*, 2014.

[14] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, 2014.