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 instead of using speech activity detection (SAD). SoftSAD improves the generalization of speech/non-speech models to unseen conditions by removing the need to make binary speech/non-speech decisions based on a threshold. Instead, softSAD explicitly integrates into the Baum-Welch statistics a speech posterior for each frame. We demonstrate the benefits of softSAD over SAD in severely mismatched conditions by evaluating a system developed for the National Institute for Standards and Technology (NIST) 2012 speaker recognition evaluation (SRE) on the channel-degraded Defense Advanced Research Projects Agency Robust Automatic Transcription of Speech speaker identification task (and vice versa). We also show that SoftSAD provides benefits over SAD in matched 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 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 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. 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 IN SPEAKER RECOGNITION

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]. Consider the impact of enrolling speakers using audio inclusive of background noise (non-speech): such a process will reduce speaker discrimination in the system, since non-speech is likely to appear similar to the system for all speakers[1]. 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 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

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).

[1] Although it is possible that speaker discrimination could appear to improve if the background noise for each speaker is unique, thus highlighting an issue with the data acquisition method.

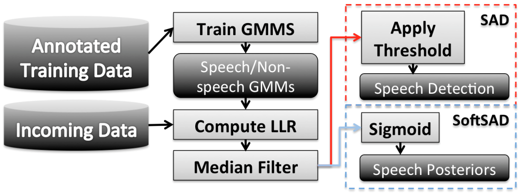


Fig. 1. The GMM-based approach to computing smoothed speech/non-speech likelihood ratios, and the subsequent SAD and softSAD processing stages.

GMMs, HMMs, neural networks and more recently, deep neural networks [1]. Neural networks are very effective under matched conditions when ample training data is available. HMMs are ideal when low-energy voiceless speech should be retained for natural language processing or user intelligibility. 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 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 [6, 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 (S) and non-speech (NS) GMM from a development data set. The likelihood ratio (LLR) of speech vs. non-speech for incoming audio frames is calculated using the trained models, and smoothed using a median filter spanning a time frame of around 400ms. Application of a threshold to these likelihood ratios results in SAD — 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 trained on 20-dimensional MFCCs (for SRE'12) or PNCCs (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,

$$post(LLR) = \frac{1}{1 + e^{-\alpha(LLR+\beta)}}. \tag{1}$$

The sigmoid parameters α and β are tuned later in the study. The zero- and first-order Baum-Welch statistics (N and F, respectively)

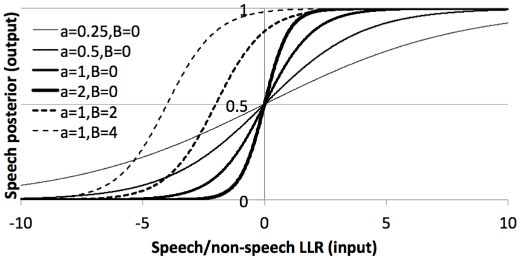


Fig. 2. The effect of varying sigmoid alpha and beta parameters on the LLR to speech posterior transformation.

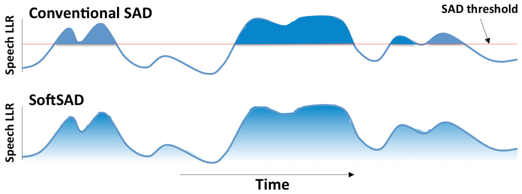


Fig. 3. SAD applies a threshold to determine speech frames which are equally weighted with respect to the SID system. The proposed softSAD method avoids a threshold and uses all speech frames weighted by their corresponding speech posteriors.

can then be calculated as:

$$N = post * \gamma \tag{2}$$

$$F = post * \gamma f \tag{3}$$

where f represents the features extracted from the audio, and γ the occupation counts for 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 to infinity and β to the SAD threshold, the same output as SAD will be obtained. By using a more tapered change in posteriors, we 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 in this work, it would be intuitive to threshold speech posteriors at a very low value (i.e., 0.05) 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

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 (train/eval): 5-8 mins / 30-200 seconds
Gender distribution: 57% female, 43% male
Evaluation language: English

RATS

Channels: 8 heavily degraded transmission channels
Noise: Push-to-talk channels with non-transmission regions
Duration (train/eval): 10-15 mins / 10 seconds
Gender distribution : 31% female, 69% male
Evaluation languages: Lev. Arabic, Dari, Farsi, Pashto, Urdu

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 features (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.

4. PROTOCOL AND SYSTEM CONFIGURATION

4.1. Tuning Protocol

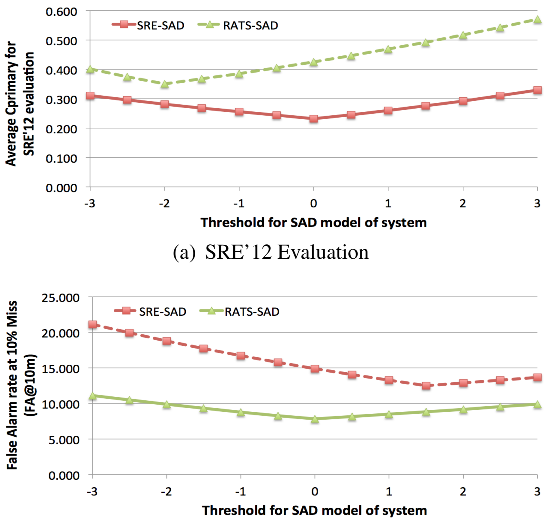
Experiments were conducted in the following manner. First, both SRE'12 and RATS systems and their corresponding SAD components were tuned independently; these are labeled as the SRE-SAD and RATS-SAD systems. The system models were then fixed before introduction of softSAD, which was tuned to optimize 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-softSAD and RATS-softSAD systems.

4.2. System Configurations

All systems are based on the i-vector/probabilistic linear discriminant analysis (PLDA) framework [7, 8]. 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. Future work will look at overcoming this drawback.

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 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 [9], the latter being an average of two distinct operating points.



(a) SRE'12 Evaluation

(b) RATS 10s-10s SID Evaluation

Fig. 4. The effect of varying the SAD threshold on SID performance for both SRE-SAD and RATS-SAD systems when evaluated on matched conditions (solid lines) and the mismatched corpora (dashed lines).

RATS System: Power-normalized cepstral coefficients (PNCC) [10] 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 [11]) for SID modeling, which extended our previous work on DCT coefficients in [12]. 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 [13] 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 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 generaliza-

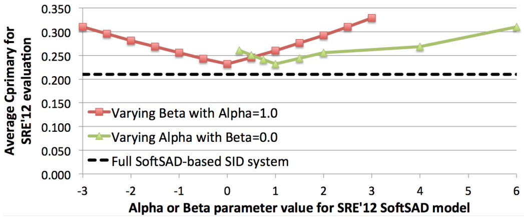


Fig. 5. The effect of varying the SRE'12 softSAD parameters on SRE'12 performance (matched condition). The β is analogous to the SAD threshold when α is infinite.

tion. 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 to highlight the improved generalization to unseen data obtained using the proposed softSAD approach.

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. Given the amount of development effort, we commence with these SAD and system models. In this section, we aim to observe the effect of changing the SAD threshold on the development set as well as the alternate corpus (for instance, the SAD threshold of the SRE'12 system on RATS SID task).

Figure 4 illustrates the difference in performance when varying the SAD speech/non-speech detection threshold on both corpora using SAD models matched to the conditions (solid line) and trained on the alternate corpora (dashed line). The best threshold for matched SAD models was 0.0, while the mismatched condition deviated from this value. Furthermore, the variation in performance on the mismatched dataset was considerably greater than in the matched case. This observation highlights the inability of SAD thresholds to generalize well to severely mismatched data.

5.2. SoftSAD Tuning

In this section, we tune the parameters on softSAD on the matched development datasets. This is done by fixing the system models (trained using a SAD threshold of 0.0) and evaluating the enroll and test audio using a variety of softSAD parameters. In both cases, maintaining a 0.0 shift and 1.0 α provided the best performance. Nonetheless, we illustrate in Figure 5 how SID performance deviates from this optimum with respect to the α and β parameters on the SRE'12 corpus. Aside from the marginal improvement observed from softSAD, it is interesting to note that changes in softSAD parameters resulted in less variation in SID performance compared to varying SAD threshold in Figure 4.

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. The former, however, may not be relevant in the current configuration due to ample system training data, but would be a valuable characteristic in low-resource training conditions. Performance after re-training is also depicted in Figure 5, with marginal gains of 5% and 10% in system performance for the SRE and RATS corpora, respectively.

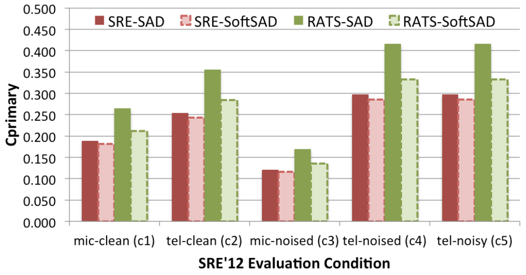


Fig. 6. Evaluation of SRE'12 extended protocol conditions using matched (SRE-SAD/softSAD) and mismatched (RATS-SAD/softSAD) systems with both SAD and softSAD approaches.

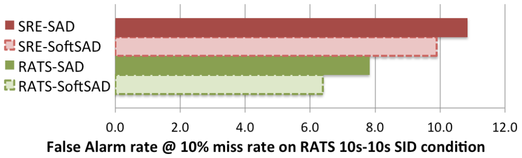


Fig. 7. Evaluation of RATS 10s-10s SID task using matched (RATS-SAD/softSAD) and mismatched (SRE-SAD/softSAD) systems.

The improvement from the red line to the full softSAD-based system (black dashed line) in Figure 5 is due to the system's awareness of soft posterior statistics. These results demonstrate that even in matched conditions, softSAD provides benefit over SAD at the cost of additional processing of frames.

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 data. Specifically, both SRE-SAD and SRE-softSAD 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 each evaluation corpus, four systems are detailed: matched and mismatched systems, each with SAD or softSAD speech modeling approaches. These results clearly show that softSAD consistently outperforms SAD in both matched and mismatched conditions. The relative improvement of softSAD over SAD in mismatched conditions was 12% and 20% for the SRE'12 and RATS evaluation, respectively — considerably greater than observed for the matched case of 5% and 10%.

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 well to severely mismatched conditions and in both matched and mismatched conditions, the proposed softSAD was a considerably more robust alternative. Future work will consider alternate methods of producing speech posteriors and whether softSAD benefits trending state-of-the-art DNN/i-vector [13] and CNN/i-vector [14] approaches to SID and language identification, since these frameworks tend to account for non-speech posteriors through explicit non-speech tri-phone state modeling.

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] 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.

[7] 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.

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

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

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

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

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

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

[14] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," in *Proc. Odyssey Workshop*, 2014.