Multistream Robust Speaker Recognition Based on Speech Intelligibility

Sridhar Krishna Nemala and Mounya Elhilali
Department of Electrical and Computer Engineering
Center for Speech and Language Processing
The Johns Hopkins University, Baltimore, MD 21218
Email: {nemala,mounya}@jhu.edu

Abstract—**Delimiting the most informative voice segments of an acoustic signal is often a crucial initial step for any speech processing system. In the current work, we propose a novel segmentation approach based on a perception-based measure of speech intelligibility. Unlike segmentation approaches based on various forms of voice-activity detection (VAD), the proposed segmentation approach exploits higher-level perceptual information about the signal intelligibility levels. This classification based on intelligibility estimates is integrated into a novel multistream framework for automatic speaker recognition task. The multistream system processes the input acoustic signal along multiple independent streams reflecting various levels of intelligibility and then fusing the decision scores from the multiple steams according to their intelligibility contribution. Our results show that the proposed multistream system achieves significant improvements both in clean and noisy conditions when compared with a baseline and a state-of-the-art voice-activity detection algorithm.**

I. Introduction

In the real world, natural speech is an amalgam of speech segments, silences and environmental and channel effects. The presence of non-speech elements in the signal hinders the performance of many speech technologies, including automatic speech and speaker recognition [1], [2], speech transmission over telephony or internet [3], noise reduction and echo cancellation [4], video conferencing [5]. This vulnerability is partly caused by speech attributes that are tailored to extracting speech elements from the signal and cannot account for quiet segments or extrinsic factors such as communication channel effects and environmental noises.

Efforts into segmenting the signal into its most informative voice components have largely employed various forms of voice-activity detection (VAD), speaker segmentation or end-point detection approaches; running the gamut from deterministic acoustic-driven to statistical methods. On one end of the spectrum, conventional VADs are based on exploring the acoustic properties of the signal (e.g. energy thresholds, zero-crossings, cepstral coefficients, etc). Most acoustic-driven approaches perform effectively (and often realtime) in controlled environments with high signal to noise ratios (SNR) and known stationary noises. However, they generally suffer a drastic drop in performance in uncontrolled environments particularly in presence of nonstationary or low SNR noises; and are susceptible to errors especially for unvoiced speech segments (e.g. fricatives and plosives). On the other end of the spectrum, efforts have focused on exploring the statistics of speech and noise sounds and employing model-based approaches and decision rules to delimit boundaries between speech- and non-speech-dominant segments. These techniques tend to be more computationally intensive, and generally fail to generalize to unseen acoustic conditions.

The present study explores a new direction for front-end pruning of the signal by augmenting the acoustic analysis with a perception-based measure of speech intelligibility. The new setup provides higher-level knowledge that complements the acoustic analysis, and is tested in the context of an automatic speaker verification (ASV) task. The proposed method offers a new way of taking advantage of the perceptual quality of the signal irrespective of its acoustics by incorporating information about the perceptual integrity of sound. Unlike statistical VAD methods, the new scheme does not require any training on the speech and noise statistics; hence making it more suitable for mixed training/testing conditions and therefore of general appeal to a wide range of applications. We propose a multistream framework based on speech intelligibility for speaker recognition. The classification based on intelligibility estimates is integrated into the multistream system by processing the test signal along multiple independent streams reflecting

978-1-4244-9848-2/11$26.00©2011 IEEE

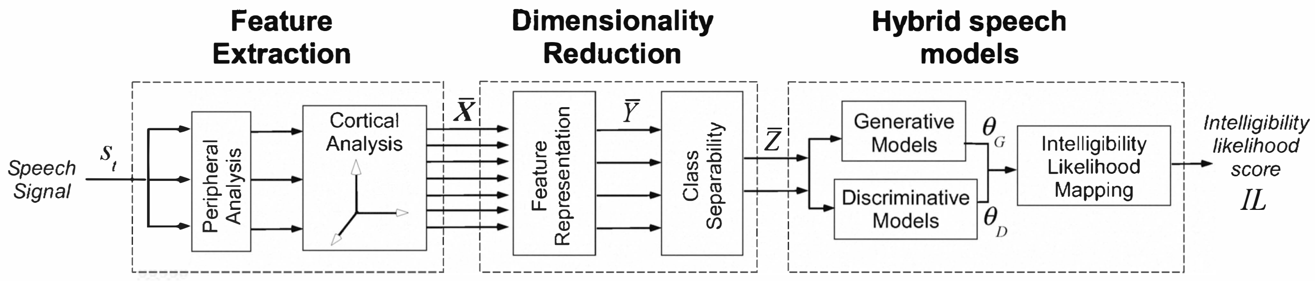

Fig. 1. Schematic of the Intelligibility Likelihood model [6]

various levels of intelligibility and fusing the scores (log-likelihood ratios) from the multiple streams according to their intelligibility contribution. The rest of the paper is organized as follows. We describe the intelligibility metric used in the proposed scheme in Section II, followed by the speaker verification system augmented with a multistream intelligibility weighting in Section III. Results of the proposed multistream system are given in Section IV followed by a discussion of the relevance of these findings and potential extensions to various speech technology systems (in Section V).

II. The Intelligibility Likelihood (IL) Model

Models of speech intelligibility aim to provide an accurate estimate of the intelligibility level of a speech signal or frame; hence replacing behavioral judgments based on human psychoacoustic testing. Most conventional intelligibility metrics; including the articulation index -AI- [7], [8], speech intelligibility index -SII- [9], speech transmission index -STI- [10], and spectro-temporal modulation index -STMI- [11], often require a reference comparison signal and compute average intelligibility scores for a given acoustic distortion or listening environment. The conventional methods rely entirely on acoustic-level analyses of the signal and compute the average intelligibility scores by a direct comparison of different attributes based on measures of spectral profile, temporal modulations, signal-to-noise ratios at different frequency bands, and spectro-temporal speech patterns. A crucial component of the proposed multistream framework is to delimit any given acoustic signal (without the need for reference signal/tempates) at the local-level (short time windows of the order of 250ms) based on higher-level perceptual information about the signal intelligibility levels. In the current study, we opt to use an Intelligibility Likelihood (IL) metric [6] which augments the acoustic-based analysis with a phonological mapping, hence bypassing the need for a reference comparison signal and enabling an assessment of the perceptual integrity of the signal over syllable length (250ms) time windows. The schematic of the IL model is shown in Fig 1.

The IL analysis starts with a biologically-inspired auditory model which mimics various stages of the mammalian auditory pathway from the periphery all the way to the primary auditory cortex. The model maps any sound into a four dimensional tensor spanning time, frequency, temporal modulations, and spectral modulations. These cortical features are then processed blockwise over short 250ms time windows, and reduced in dimensionality using higher-order singular value decomposition [12] followed by modified linear discriminant analysis [13]. Finally, a hybrid generative/discriminative statistical model of highly-intelligible and none-intelligible speech classes is built. An Intelligibility Likelihood (IL) score is then obtained by fusion of the likelihood measures from the statistical models using a two-layer feed-forward neural network. Full details about the implementation of the IL model are provided in [6].

For the IL model used here, the highly-intelligible speech class is learned from approximately one hour of *interview-microphone* speech data taken from NIST 2008 speaker recognition evaluation [14]. To train the none-intelligible speech class, the same data but with white noise added at -20dB SNR is used. For the data and results presented in this work, we use only the discriminative statistical model component from the full model to compute the IL metric as it significantly reduces the computational complexity of the full model[1]. Given an acoustic signal, a stream of IL estimates from the model enables local-level tracking of transitory changes in intelligibility. An example of how the IL metric tracks changes in the signal intelligibility is shown in Fig 2.

[1] Note that for the ASV task explored in this paper, it is not required to compute an exact intelligibility score (0-100%) like in the original IL model

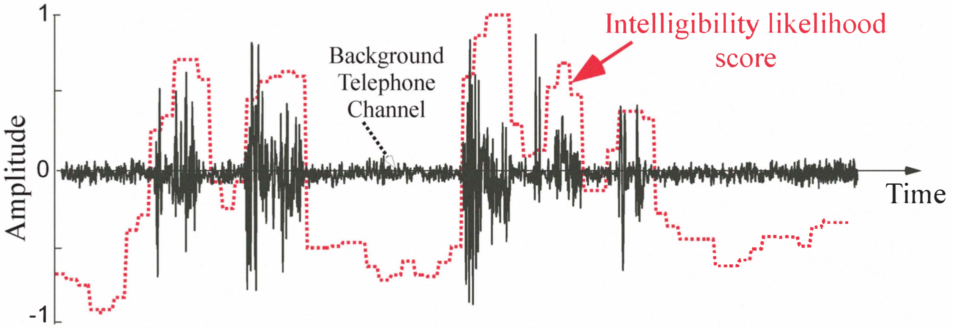

Fig. 2. Illustration of tracking of transitory changes in signal intelligibility. The test utterance is a noisy telephone conversation clip taken from the NIST 2008 speaker recognition evaluation

III. Speaker Recognition Setup

Gaussian Mixture Model (GMM) based speaker verification systems form the state-of-the-art and have been shown to give excellent performance in all the recent NIST speaker recognition evaluations (SREs). In GMM-based speaker verification setup, a speaker-independent Universal Background Model (UBM) is first trained with data gathered from a large number of speakers [15]. The UBM represents speaker-independent distribution of the feature vectors. When enrolling new speakers to the system, models for the target speakers are obtained by *maximum a posteriori* (MAP) adaptation of the UBM. In the verification stage, a match score is computed in the form of a log likelihood ratio - which essentially is a measure of the differences between target speaker model and the speaker-independent UBM in generating the test speaker observations (feature vectors).

In our UBM-GMM based speaker recognition system, we trained the UBM with data obtained from a set of 325 speakers. The data is sampled from the NIST 2008 SRE [14] training corpus. In the UBM training, a total of 1024 mixtures and 15 expectation-maximization iterations for mixture split are used. A total of 137 target speaker models are obtained by MAP adaptation of the UBM. MIT Lincoln Lab GMM toolkit is used for the UBM-GMM training. For the verification task, we focus in particular on condition 4 even though there are eight common conditions listed in the NIST 2008 SRE [14]. This condition represents a *mismatch* scenario, where the trials involve interview (microphone) training speech and telephone test speech, and thereby provides with an ideal case for studying the robustness aspect of the speaker recognition system. For the front-end acoustic features, standard 19 Mel Frequency Cepstral Coefficients (MFCC) along with their first and second order temporal derivatives are used. A 25ms analysis window and 10ms time shift is used in the feature vector computation. In addition, utterance level mean and variance normalization is employed.

A. Multistream framework

The motivation for choosing an incongruent or mismatched training/testing condition for the ASV setup is to allow a more realistic testing of the system. In this context, it is very important to focus on more informative portions in the noisy test signal. Towards that end, we propose a speech intelligibility based multistream framework for the speaker recognition task. A schematic of the proposed approach is shown in Fig 3. During verification, the test utterance is analyzed using the IL model (described in Section 2). Based on the perceptual quality of the signal (which has been shown to correlate highly with human speech intelligibility judgements [6]), the input feature stream is segregated into a number of different streams. This is achieved by (i) computing the local-level IL estimates from the IL model for the entire test utterance; (ii) clustering the IL scores into desired number of clusters or feature streams (k-means clustering is used); (iii) generating multiple streams corresponding to the IL clustering, eg. high, high-medium, medium, medium-low, and low intelligible feature streams. In the current study, we found that a choice of 5 streams was optimal. In the final stage, the decision scores or log likelihood ratios are computed independently for each of the multiple streams, and stream fusion at the scores level is performed based on appropriate weighting. This weighting can be directly related to the respective streams average intelligibility estimates, and is empirically found to be optimal with the following weight values: [0.45 0.25 0.15 0.1 0.05][2]. Note

[2] It is also possible to learn these weights using data driven techniques - Fusion and Calibration toolkit, http://www.dsp.sun.ac.za/~nbrummer/focal

978-1-4244-9848-2/11$26.00©2011 IEEE

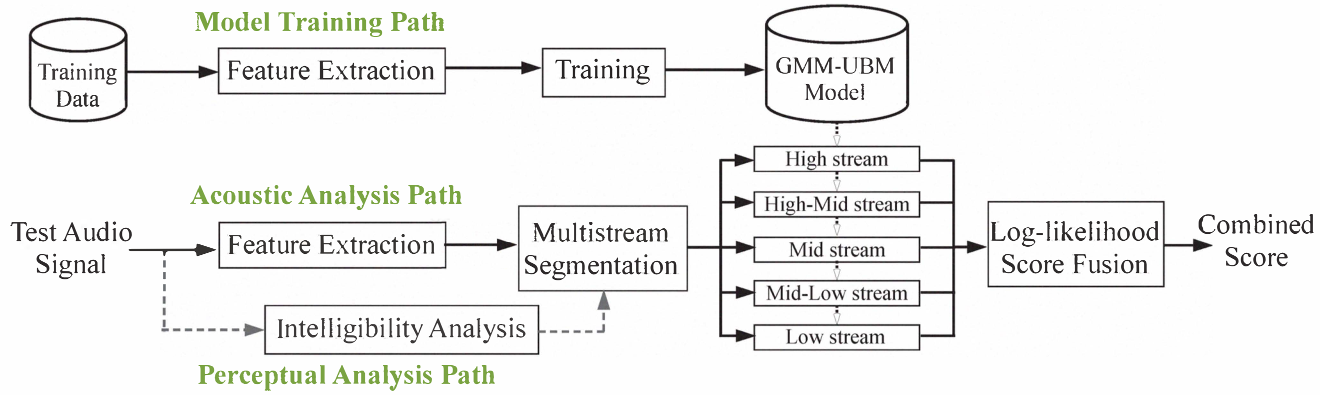

Fig. 3. Schematic of the Intelligibility Likelihood (IL) based multistream speaker recognition system

that the optimal weight distribution reflects the higher importance given to high intelligible streams.

IV. Experiments and Results

We test the proposed multistream framework for the mismatch condition 4 (interview train and telephone test) in the NIST 2008 SRE. A subset of 137 train speakers is chosen to train the target speaker models and a set of 1073 test trials are used to evaluate the verification performance. In the first set of experiments, we evaluate the verification performance for a baseline system (with no VAD component); as well as for each of the 5 intelligibility-streams in the multistream system. The results in Equal Error Rates (EER) are shown in Table 1. It can be seen that the performance of different streams correlates highly with the intelligibility level of the individual streams, further validating the proposed intelligibility based segregation of feature streams. The multistream fusion at the decision score level achieved the best verification performance 21.56% EER. Note that the combination also improves over the best single stream performance in the multiple streams.

In the second set of experiments, we further corrupt the test signal with white noise added at different Signal-to-Noise Ratios (SNR). This addition of external sources of noise on top of the channel mismatch condition represents a scenario which the speaker recognition systems are likely to deal with in any real world application. Note that in both set of experiments, we do not assume external sources of information (or segmentation) coming from Automatic Speech Recognition (ASR) systems. This assumption is particularly valid as the ASR systems also suffer a drastic drop in performance in mismatch and noisy conditions [16]. Table 2 shows the performance with the baseline system, performance obtained with pruning using state-of-the-art ETSI VAD [17], and performance with the proposed multistream fusion system. While VAD based pruning improves over the baseline performance, note that the multistream system performs significantly better in all the conditions - an average EER reduction of 26% and 18% respectively, over baseline system and ETSI VAD based pruning[3].

TABLE I
Speaker verification performance for the interview train and telephone test mismatch condition (results are in Equal Error Rate - EER in percentage)

Stream Description	Performance (EER)
Baseline	31.51
High Intelligible	24.30
High-Medium Intelligible	26.46
Medium Intelligible	27.60
Medium-Low Intelligible	43.13
Low Intelligible	48.03
Multistream Combination	21.56

V. Discussion

We propose a multistream framework based on speech intelligibility for speaker recognition. Unlike current segmentation approaches, the proposed method focuses on labeling the test signal according to the perceptual integrity of its content (i.e. how intelligible it would sound to a human listener irrespective of its physical acoustics). This classification based on intelligibility estimates is integrated into the ASV system by processing the test signal along multiple independent streams

[3] In our experiments, other conventional VADs based on energy thresholds, zero crossings, spectral/cepstral measures etc. did not improve results over the ETSI VAD

TABLE II
SPEAKER VERIFICATION PERFORMANCE FOR THE INTERVIEW TRAIN AND TELEPHONE TEST MISMATCH CONDITION. IN THIS THE TEST SIGNAL IS FURTHER CORRUPTED BY ADDITIVE WHITE NOISE (RESULTS ARE IN EQUAL ERROR RATE - EER IN PERCENTAGE)

Noise	Performance (EER)		
Condition	Baseline	ETSI VAD	Multistream
Clean ∞ dB	31.51	27.81	21.56
white noise, 30dB	39.19	35.25	27.45
white noise, 20dB	42.74	38.72	33.08
white noise, 10dB	45.13	41.21	35.29

reflecting various levels of intelligibility and fusing the scores (log-likelihood ratios) from the multiple streams according to their intelligibility contribution. The proposed system is tested in a realistic setting where the test signal comes from a *mismatch* condition (both channel and environmental). We show significant improvements over the baseline system as well as a state-of-the-art VAD-based pruning approach. How the improvements would hold in conjunction with the channel and noise compensation techniques remain to be seen.

In the multistream framework, we show that the performance of different streams correlates highly with the intelligibility level of the individual streams which further validates the proposed segregation of feature streams. Since the local-level speech intelligibility estimates can be looked at as additional higher level knowledge sources, we speculate the proposed multistream system would be applicable even in *matched* testing conditions - the benefits may be limited, however. Further, a similar multistream front-end is not limited to ASV tasks, but can be applied to other speech processing applications such as ASR, noise reduction or speech enhancement, and telephone/internet speech transmission.

ACKNOWLEDGMENT

This research is partly supported by NSF CAREER grant IIS-0846112, AFOSR grant FA9550-09-1-0234 and NIH grant 5R01AG036424, and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

REFERENCES

[1] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," *Speech Comm*, vol. 3, p. 261276, 2003.

[2] T. N.-L. V. Hautamki, M. Tuononen and P. Frnti, "Improving speaker verification by periodicity based voice activity detection," *Proc. 12th International Conference on Speech and Computer (SPECOM)*, 2007.

[3] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. Venkatesha Prasad, and V. Gaurav, "Vad techniques for real-time speech transmission on the internet," in *Proc. High Speed Networks and Multimedia Communications 5th IEEE Int. Conf*, 2002, pp. 46–50.

[4] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Comm*, vol. 16, pp. 245–254, 1995.

[5] D. M. B. Lee, "Spectral entropy-based voice activity detector for videoconfencing systems," in *proceedings of Interspeech*, 2010.

[6] S. K. Nemala and M. Elhilali, "A joint acoustic and phonological approach to speech intelligibility assessment," in *Proc. IEEE Int Acoustics Speech and Signal Processing (ICASSP) Conf*, 2010, pp. 4742–4745.

[7] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J Acoust Soc Am*, vol. 17, no. 1, p. 103, 1945.

[8] ANSI-S3.5-1969-R1978, "Methods for the calculation of the articulation index," 1969.

[9] ANSI-S3.5-1997-R2007, "Methods for calculation of the speech intelligibility index," 1997.

[10] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J Acoust Soc Am*, vol. 67, pp. 318–326, 1979.

[11] M. Elhilali, T. Chi, and S. A.Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Comm*, vol. 41, pp. 331–348, 2003.

[12] B. D. M. L. De Lathauwer and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Applicat.*, vol. 21, pp. 1253–1278, 2000.

[13] S. Chen and D. Li, "Modified linear discriminant analysis," *Pattern Recog*, vol. 38, pp. 441–443, 2005.

[14] "NIST 2008 Speaker Recognition Evaluation," http://www.nist.gov/speech/tests/sre/2008.

[15] R. D., Quatieri, and T. D. R, "Speaker verification using adapted gaussian mixture models," *Digital Signal Process.*, 2000.

[16] A. Waibel and K. Lee, Eds., *Readings in speech recognition.* SF: Morgan Kaufmann Pub. Inc., 1990.

[17] ETSI, "Voice activity detector (vad) for adaptive multi-rate (amr) speech traffic channels," *Sophia Antipolis, France, ETSI EN 301 708 Rec.*, 1999.

978-1-4244-9848-2/11$26.00©2011 IEEE