MULTILEVEL SPEECH INTELLIGIBILITY FOR ROBUST SPEAKER RECOGNITION

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
{nemala,mounya}@jhu.edu

ABSTRACT

In the real world, natural conversational speech is an amalgam of speech segments, silences and environmental/background and channel effects. Labeling the different regions of an acoustic signal according to their *information* levels would greatly benefit all automatic speech processing tasks. 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 parsing approach exploits higher-level perceptual information about signal intelligibility levels. This labeling information is integrated into a novel multilevel framework for automatic speaker recognition task. The system processes the input acoustic signal along 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 system achieves significant improvements over standard baseline and VAD-based approaches, and attains a performance similar to the one obtained with oracle speech segmentation information.

Index Terms— Speech intelligibility, Voice-activity detection, Speaker recognition, Noise robustness

1. INTRODUCTION

With the advent of E-commerce technology, the importance of non-intrusive and highly reliable methods for personal authentication has been growing rapidly. Voice prints being the most natural form of communication, and being already used widely in spoken dialog systems, have significant advantage over other biometrics such as retina scans, face, and finger prints. Voice prints as biometric also have tremendous potential in forensic and military applications. However despite significant advances in Automatic Speaker Verification/Recognition (ASV) over the last two decades, the real-world application of ASV technology still faces tremendous challenges. The presence of corrupted speech (at various degrees and in a variety of ways) as well as non-speech segments hinders the performance of ASV systems, particularly in the context of conversational speech biometrics [1].

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. Most of the existing approaches suffer a significant drop in performance in uncontrolled/noisy environments and unseen acoustic conditions. The present study explores a new direction for front-end pruning of the signal with a perception-based measure of speech intelligibility. The proposed method offers a new way of taking advantage of the perceptual quality of the signal irrespective of its acoustic environment by incorporating information about the perceptual integrity of the sound. We propose a multilevel system for speaker recognition that makes an additional use of the information about speech intelligibility levels present in the input acoustic signal. The labeling information based on time-varying intelligibility estimates is integrated into the multilevel system by processing the test signal along multiple independent streams reflecting 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 2, followed by the speaker verification system augmented with a multistream intelligibility weighting in Section 3. Results of the proposed multilevel system are given in Section 4 followed by a discussion of the relevance of these findings and potential extensions to various speech processing systems (in Section 5).

2. THE INTELLIGIBILITY LIKELIHOOD MODEL

In any conversational speech biometrics application, it is very important to identify and make use of the different regions in the test signal that contain different levels of *information*. In this work, we propose to use information based on speech

This research is partly supported by NSF CAREER grant IIS-0846112, NIH grant 1R01AG036424 and AFOSR grant FA9550-09-1-0234, 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.

intelligibility levels. Most conventional intelligibility metrics; including the articulation index -AI- [2, 3], speech intelligibility index -SII- [4], speech transmission index -STI- [5], and spectro-temporal modulation index -STMI- [6], often require a reference comparison signal and compute average intelligibility scores for a given acoustic distortion or listening environment. A crucial component of the proposed multilevel system framework is to label any given acoustic signal, without the need for reference signal, 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 use a variation of the Intelligibility Likelihood (IL) model that was originally proposed in [7] to enable the assessment of the perceptual integrity of any given signal over syllable length (250ms) time windows.

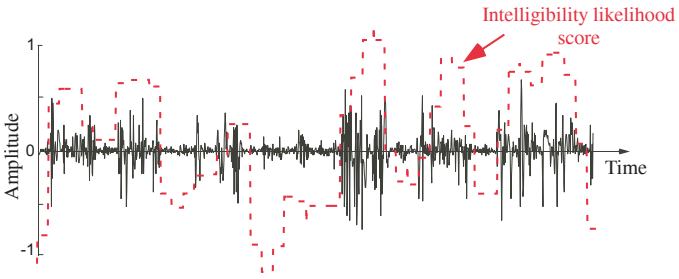

Fig. 1. Illustration of tracking of transitory changes in signal intelligibility levels. The speech utterance is an interview conversation clip taken from the NIST 2008 SRE.

The IL analysis starts with a biologically-inspired auditory model which mimics various stages of the mammalian auditory pathway. The auditory model contains two basic stages: an early stage that maps the one-dimensional acoustic signal to a time-frequency representation (auditory spectrogram), and a subsequent central stage that analyzes the auditory spectrogram to estimate its modulation profile along spectral and temporal dimensions using a bank of modulation-selective filters. The model maps any sound into a four dimensional cortical representation spanning time, frequency, temporal modulations (rates), and spectral modulations (scales). Full details of the auditory model are given in [8]. In the current implementation, 90 frequency channels placed uniformly over log-frequency scale in the range of 300-3400 Hz, rate filters selective to $[4, 8, 16, 30, 50]$ Hz, and scale filters selective to $[0.25, 0.5, 1, 2, 4]$ Cycles/Octave are used. The cortical features are then processed blockwise integrating over short 250ms time windows, and reduced in dimensionality using higher-order singular value decomposition [9] to 72 dimensions (6 x 4 x 3 in frequency, rate, scale subspaces, respectively). In this reduced dimensional space, a Support Vector Machine (SVM) is trained to discriminate high-intelligible and low-intelligible speech classes. Given any reduced test cortical feature representation, an IL score is computed from the distance between the feature vector and the separating hyperplane in the SVM. Full details about the implementation of the IL model are provided in [7]. For the IL model used here, the high-intelligible features are computed from approximately one hour of TIMIT speech database [10]. For the low-intelligible speech class, the same data but with white noise added at -10dB Signal-to-Noise Ratio (SNR) is used. 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 1.

3. 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 on matched-channel condition 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 [11]. The UBM represents speaker-independent distribution of the feature vectors. When enrolling new speakers into 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 402 speakers. The data is sampled from the NIST 2008 SRE [12] training corpus. In the UBM training, a total of 1024 mixtures and 15 expectation-maximization iterations for mixture split are used. A total of 85 target speaker models are obtained by MAP adaptation of the UBM. MIT Lincoln Lab GMM toolkit is used for the UBM-GMM training.

3.1. Multilevel system framework

We propose a multilevel system framework based on speech intelligibility for the speaker recognition task. A schematic of the proposed system is shown in Fig 2. During verification, the test utterance is analyzed using the IL model (described in Section 2). Based on the perceptual quality of the signal that has been shown to correlate highly with human speech intelligibility judgements [7], the input feature stream is segregated into a number of different streams. This is achieved by **(i)** computing IL estimates over short time scales for the entire test utterance; **(ii)** clustering the array of IL scores in an unsupervised way into desired number of clusters or feature streams (k-means clustering is used); **(iii)** generating multiple feature streams corresponding to the IL score clustering, e.g., high, medium, and low intelligible level feature streams. In the current study, for microphone interview recordings, we

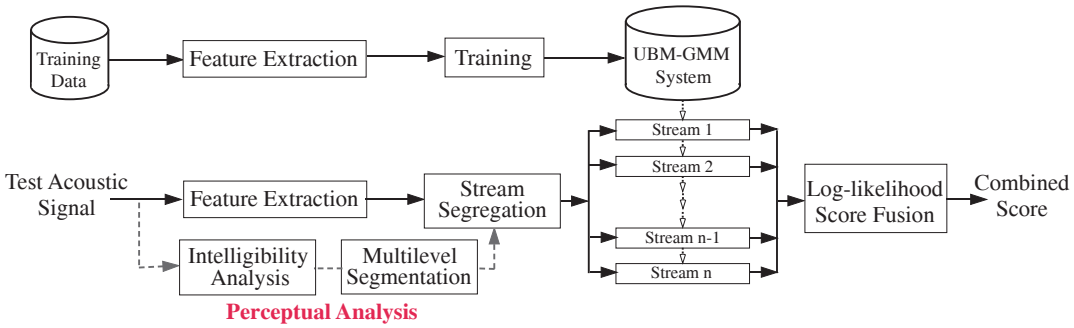

Fig. 2. Schematic of the multilevel system based on speech intelligibility for speaker recognition

found that a choice of 3 streams is optimal. For telephone channel recordings in the NIST SRE data, we observed that a choice of 5 streams is optimal (results not presented here). The IL score clustering results in segregation of the feature streams in an adaptive manner, for e.g., a larger percentage of frames are assigned to lower intelligibility streams when the test signal is relatively noisy or corrupted. 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: [0.75 0.25 0][1]. Note that the optimal weight distribution reflects the higher importance given to high intelligible streams.

4. EXPERIMENTS AND RESULTS

For the verification task, we focus in particular on condition 2 even though there are eight common conditions listed in the NIST 2008 SRE [12]. In this condition, both the train and test trials involve interview conversations from the *same* microphone. We specifically choose this condition, since the recognition setup cannot take advantage of factor analysis techniques [13] that address various channel mismatch scenarios present in the standard NIST SREs. A subset of 85 train speakers is chosen to train the target speaker models and an independent set of 500 test trials (impostor and genuine trials are 169 and 331, respectively) taken from NIST 2008 SRE is used to evaluate the verification performance. For the front-end acoustic features, standard 19 Mel Frequency Cepstral Coefficients (MFCC) along with their first and second order temporal derivatives are used. In addition, utterance level mean and variance normalization is employed.

In the first set of experiments, we evaluate the verification performance for a baseline system that uses the entire test signal without the use of intelligibility level information; as well as for each of the 3 streams with high, medium, and low intelligibility levels in the multilevel 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. The multistream fusion at the decision score level achieved the best verification performance of 3.3% EER. Notice that the combination also improves over the best single stream performance.

Table 1. Speaker verification performance – results are in Equal Error Rate (EER) in percentage

Stream Intelligibility Level	ASV Performance (in EER)
Baseline	4.5
1 (high)	3.6
2 (medium)	5.3
3 (low)	26.6
Multistream Combination	**3.3**

In the second set of experiments, we corrupt the test signal with two different real-world noise types added at Signal-to-Noise Ratio (SNR) levels of 30dB, 20dB, and 10dB. The noise types chosen are non-stationary subway and street noise (taken from Aurora database [14]). Table 2 shows the performance obtained with **(i)** the baseline system **(ii)** with pruning using state-of-the-art ETSI VAD [15] **(iii)** with the proposed multilevel system based on speech intelligibility information **(iv)** with pruning using the segmentation coming from an Oracle Automatic Speech Recognition (ASR) system - note that the oracle ASR system uses *clean* signal to obtain the segmentation, but in the real world test scenario, the availability of clean reference signal is not practical[2]. VAD based pruning improves over baseline performance especially in clean and high SNR conditions. While VAD based approach gives marginal improvement over baseline, the multilevel system performs significantly better in all the conditions - an average EER reduction of 33% and 29% respectively, over baseline system and ETSI VAD based pruning[3]. Note that proposed multilevel system performs comparable to the system that uses Oracle ASR segmentation information.

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

[2]Further, ASR systems suffer a significant drop in performance in mismatch, noisy, and unseen acoustic conditions

[3]Other conventional VADs based on energy thresholds, zero crossings, and spectral/cepstral measures did not improve results over the ETSI VAD

Table 2. Speaker verification performance where the test signal is corrupted by nonstationary additive noise – results are in Equal Error Rate (EER), in percentage

Noise Type	SNR (in dB)	ASV Performance (EER)			
		Baseline	ETSI VAD	Multilevel System	Oracle ASR
Clean	∞	4.5	4.1	**3.3**	3.5
Street	30	4.8	4.2	**3.5**	3.9
	20	7.7	7.0	**3.9**	4.1
	10	13.6	13.4	8.7	**8.3**
Subway	30	8.9	7.7	**3.5**	3.6
	20	13.4	12.6	8.9	**8.7**
	10	22.1	21.8	**18.3**	**18.3**

5. DISCUSSION

We propose a novel approach for labeling any given acoustic signal at short time scales with *information* levels based on a perception-based measure of speech intelligibility. This labeling information is integrated into an ASV system by processing the test signal along multiple independent streams reflecting various levels of intelligibility and fusing the scores (log-likelihood ratios) from the multiple streams according to their intelligibility contribution. Using the proposed multilevel system, we show significant improvements over standard baseline and VAD based approaches, and achieve performance close to the one obtained with oracle segmentation information. Note the IL model used for intelligibility analysis is neither trained on the train/test databases nor on the noise types being tested. It is an *independent* measure of intelligibility that closely matches human listeners' judgment of speech integrity [7]; and assumes no prior knowledge of the dataset or noise background.

In the multilevel system, we show that the performance of different streams correlates highly with the intelligibility level of the individual streams which further validates the proposed labeling and segregation of feature streams. Since the speech intelligibility level estimates can be looked at as additional higher-level knowledge sources and provides complimentary information, we speculate the proposed multilevel system would be applicable even in the context of complex large scale ASV systems that include techniques like factor analysis. Further, a similar multilevel front-end is not limited to ASV tasks, but can be applied to other speech processing applications such as speech recognition, speech transmission over telephony or internet, noise reduction and echo cancellation, and video conferencing.

6. REFERENCES

[1] H. Beigi, *Fundamentals of Speaker Recognition*, Springer, 2011.

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

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

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

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

[6] 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.

[7] 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.

[8] P. Ru T. Chi and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J Acoust Soc Am*, vol. 118, pp. 887–906, 2005.

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

[10] "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, LDC93S1*, 1993.

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

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

[13] T. Kinnunen and H. Lib, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Comm*, vol. 52, pp. 12–40, 2010.

[14] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, September 18-20. 2000.

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