A JOINT ACOUSTIC AND PHONOLOGICAL APPROACH TO SPEECH INTELLIGIBILITY ASSESSMENT

Sridhar Krishna Nemala and Mounya Elhilali

Department of Electrical and Computer Engineering, Johns Hopkins University
{ nemala,mounya@jhu.edu}

ABSTRACT

While current models of speech intelligibility rely on intricate acoustic analyses of speech attributes, they are limited by the lack of any linguistic information; hence failing to capture natural variability of speech sounds and confining their applicability to average intelligibility assessments. Another important limitation is that the existing models rely on the use of reference clean speech templates (or average profiles). In this work, we propose a novel approach to speech intelligibility by combining a biologically-inspired acoustic analysis of peripheral and cortical processing with phonological statistical models of speech using a hybrid GMM-SVM system. The model results in a novel scheme for speech intelligibility assessment without the use of reference clean speech templates, and the model predictions strongly correlate with scores obtained from human listeners under a variety of realistic listening environments. We further show that the proposed model enables local level tracking of intelligibility and also generalizes well to multiple speech corpora.

Index Terms— Speech intelligibility, spectro-temporal, psychoacoustic, statistical model, hybrid GMM-SVM

1. INTRODUCTION

Speech communication in noise poses a common challenge to both engineering systems and the human brain alike. While a nontrivial task, objectively predicting the intelligibility of speech under various distortions relies on a number of physical properties of either the signal or communication channel. The common methods; including the articulation index -AI- [1, 2], speech transmission index -STI- [3], speech intelligibility index -SII- [4], and spectro-temporal modulation index -STMI- [5], perform an acoustic-level analysis based on measures of spectral profile, temporal modulations, signal-to-noise levels at different frequency bands, and spectro-temporal speech patterns; all speech attributes closely correlated with average intelligibility. While these measures made notable progress in accurately predicting mean intelligibility under various listening and transmission conditions, they fall short in two respects: **(a)** they rely on reference clean speech templates, special test signals, or generic speech references which fail to capture the inherent variability in natural speech; **(b)** they are global measures that can only predict mean intelligibility scores for a given acoustic distortion or listening environment.

By contrast, speech processing in the biological auditory system has the advantage of employing both a robust front-end pathway spanning the outer ear all the way to auditory cortex; as well as phonological and syntactic knowledge which complements the acoustic-level analysis. In the present work, we propose a model inspired from this biological scheme which comprises both an acoustic-level analysis motivated by the processing in the auditory pathway, mediated by a phonological mapping employing hybrid generative and discriminative models. The model results in a novel scheme for speech intelligibility assessment without the use of reference clean speech templates. The schematic of the proposed model is shown in Fig 1. The model evaluates intelligibility likelihoods at a local-level (phonemic or subsyllabic), and we show that the model predictions strongly correlate with scores obtained from human listeners under a variety of realistic listening environments (Section 3.1). We further show that the proposed model enables local level tracking of intelligibility (Section 3.2) and also generalizes well to multiple speech corpora (Section 3.3).

2. INTELLIGIBILITY LIKELIHOOD (IL) MODEL

2.1. The Auditory Model

The proposed Intelligibility Likelihood (IL) method starts with a biologically-inspired auditory model which mimics the various stages taking place along the auditory pathway from the periphery all the way to the primary auditory cortex. The model consists of two basic stages: An early stage that mimics cochlear and mid-brain processing maps the one-dimensional acoustic stimulus to a time-frequency representation (auditory spectrogram), via a sequence of constant-Q cochlear filters, hair cell transduction model and lateral inhi-

This research was supported in part by NSF grant (IIS-0846112), Air Force Office for Scientific Research grant (FA9550-09-1-0234) and Applied Physics Laboratory agreement (SG204D21).

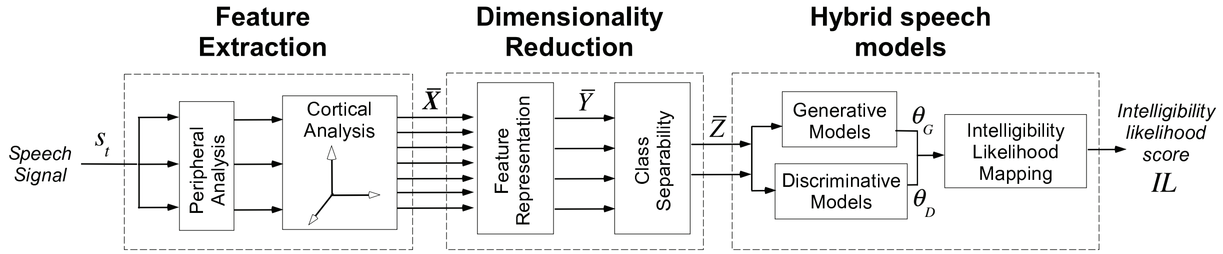

Fig. 1. Schematic of the Intelligibility Likelihood model

bition network. A subsequent central stage analyzes the auditory spectrogram to estimate the modulation profile along the spectral and temporal dimensions. This analysis is performed using a bank of modulation-selective filters, mimicking the array of feature selective filters (called Spectro-Temporal Receptive Fields, STRF) observed at the level of the mammalian primary auditory cortex (A1). The cortical analysis is mathematically equivalent to a two-dimensional affine wavelet transform of the auditory spectrogram. Full details of the auditory model are given in [5] and [6].

2.2. Dimensionality Reduction

The output of auditory model is a multidimensional array in which modulations are presented along the four dimensions of time, frequency, temporal modulations (rate), and spectral modulation (scale). The model output is processed blockwise with window size 125 ms and shift size 100 ms. Each block is then time-averaged, yielding a three-mode tensor ($\bar{\boldsymbol{X}}$) with each element representing the overall modulations at a corresponding scale, rate, and frequency. Sufficient number of filters in each mode are required to obtain a good resolution. However as a result, the dimensionality of the feature space is very high ($>$ 7000) rendering training of any statistical model impractical. We address the issue of high dimensionality using a multi-linear dimensionality reduction method followed by discriminant analysis.

2.2.1. Higher-Order Singular Value Decomposition

We employ a multi-linear dimensionality reduction procedure based on Higher-Order Singular Value Decomposition (HOSVD) [7]. HOSVD is a generalization of singular value decomposition to tensors, where every mode-n tensor D can be written as the product $D = S \times_1 U^{(1)} \times_2 U^{(2)} ... \times_N U^{(N)}$. $U^{(n)}$ is a unitary matrix containing left singular vectors of the mode-n unfolding of the tensor D, and S is a mode-n tensor which has the properties of all-orthogonality and ordering.

In the proposed approach, the tensor D contains a stack of 3-mode cortical output tensors ($\bar{\boldsymbol{X}}^{\boldsymbol{i}}$) from a set of training samples (300 sec of clean TIMIT speech data and the same 300 sec of speech data masked with -20 dB white noise). D is decomposed as

$$D = S \times_1 U_{scale} \times_2 U_{rate} \times_3 U_{frequency} \times_4 U_{samples} \quad (1)$$

in which U_{scale}, U_{rate}, $U_{frequency}$ are ordered orthonormal matrices containing the respective subspace singular vectors. Each singular matrix is then truncated by setting a threshold so as retain only a desired number of principal components (PC) in the corresponding subspace. The threshold for the number of PCs is determined to be 4 for scale, 5 for rate, 7 for frequency subspace, and these PCs preserve greater than 90% variance in their associated subspace. Given a 3-mode cortical tensor ($\bar{\boldsymbol{X}}$), it is projected onto the truncated orthonormal matrices U'_{scale}, U'_{rate}, $U'_{frequency}$ by

$$\bar{Y} = \bar{\boldsymbol{X}} \times_1 U'^{\mathrm{T}}_{scale} \times_2 U'^{\mathrm{T}}_{rate} \times_3 U'^{\mathrm{T}}_{frequency} \quad (2)$$

The result $\bar{Y}$ is then vectorized yielding a feature vector of dimension 140.

2.2.2. Modified Linear Discriminant Analysis

In order to identify the most discriminating subspace between highly and non-intelligible speech classes and further reduce the feature space dimension, we employ Linear Discriminant Analysis (LDA). Classic Fisher LDA is limited to only *one* optimal projection for a two-class problem due to the rank limitations of its between-class scatter matrix. Instead, a modified LDA (MLDA) method offers a generalization of the FLDA and overcomes the rank limitation by redefining the scatter matrix [8]. We employ MLDA to map the vector $\bar{Y}$ into a reduced feature vector $\bar{Z}$ of dimension 60.

2.3. The Statistical Model

Highly (respectively, non-) intelligible speech samples processed through the acoustic analysis and dimensionality reduction define a probabilistic distribution that delimits the natural variability in clean (respectively, noisy) speech ensembles. We estimate these densities with two classes of statistical models: one that models the underlying distribution of each class (using GMMs), while the other that maximizes the separability between these two-class distributions (using SVM). The outputs of both models are then mapped onto an *intelligibility likelihood* score via a neural network.

2.3.1. *The Generative Model*

A first Gaussian Mixture Model (GMM^H) is trained on the features derived from highly intelligible speech samples from 3000 sec of clean TIMIT speech data. We observe that the GMM training results in implicit clustering of sub-lexical units eventhough no explicit phonological segmentation is performed during the acoustic analysis[1]. A second GMM^N, counterpart to the first one, is trained on features derived from non-intelligible (noisy) speech samples. The training of GMM^N uses speech data masked at -20dB white noise. For a test input $\bar{Z}$, both GMMs produce likelihood estimates which are then converted into class conditional posterior probabilities by the following procedure: 1) Rank all (highly and non-intelligible) training samples by their likelihood scores 2) Divide the samples into n subsets of equal size bins 3) Given a test feature $\bar{Z}$, based on its likelihood with respect to each class, place $\bar{Z}$ in the corresponding bin 4) The class conditional posterior probability is given by the fraction of true positives in the bin, i.e. the fraction of training samples in the bin that actually belong to the given class. The posteriors from both the highly and non-intelligibility speech classes are then combined together as a posterior ratio, given by $\mathcal{R} = \frac{P(H|\bar{Z})}{P(H|\bar{Z})+P(N|\bar{Z})}$, where $P(H|\bar{Z})$ and $P(N|\bar{Z})$ are the class conditional posterior probability estimates with respect to GMM^H and GMM^N respectively. We use $P(H|\bar{Z})$ as the output from the generative model, as both the $P(H|\bar{Z})$ and $\mathcal{R}$ yield equivalent results for the noise conditions tested in this paper.

2.3.2. *The Discriminative Model*

A Support Vector Machine (SVM) is trained on the same highly and non-intelligible speech ensembles (used for the GMMs training) in order to define an optimal decision boundary that maximizes the separation between the two classes. We use radial Basis functions for the SVM kernel, and define the highly intelligible class as positive samples. Given a test feature $\bar{Z}$, the distance between $\bar{Z}$ and the separating hyperplane is taken as the output from the discriminative model.

2.3.3. *The Intelligibility Likelihood (IL) Score*

The outputs of the generative and discriminative models often contain useful complementary information about the coordinates of the test vector in the feature space. In order to map these outputs onto an intelligibility likelihood (IL) score, we use a two-layer feed-forward neural network with five hidden neurons trained on data obtained from human listeners tested in different listening environments (described in section 3.1). The data is divided into 90% training and 10% test sets. Evaluation of the model is done with 10-fold cross validation.

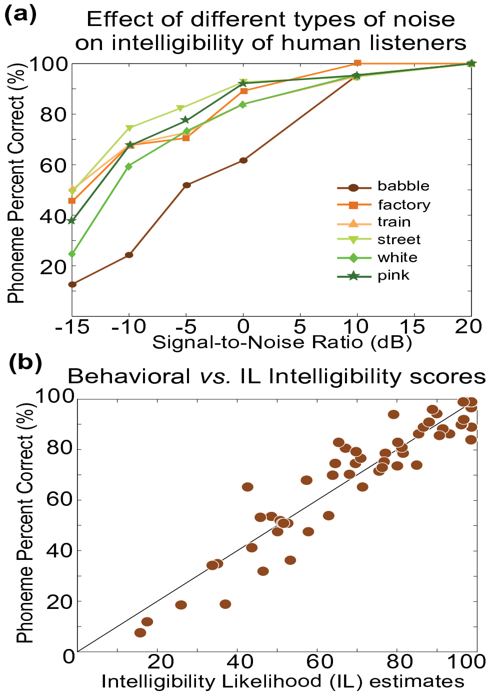

Fig. 2. Listeners performance for different noise conditions, and the correspondence between the behavioral data and IL scores

3. RESULTS

3.1. Comparing IL scores to human perception

To validate the proposed method, IL scores are compared to human intelligibility scores from six listeners using speech samples contaminated by different types of additive noise (taken from NOISEX-92 database) at different SNR levels. The noise types include white, babble, train, street, factory1, and pink, and are tested in a range between [-15,20]dB. The speech samples are a set nonsense sentences of a male speaker, each constructed from five randomly chosen monosyllabic words [5] (data provided from the SouthWest Research Institute). Each word consists of three phonemes with approximately balanced presentations of vowels and consonants. A count of correct phonemes reported is then averaged across all subjects and all test material for each noise condition. The percent correct recognition scores from the listening experiments are given in Fig 2a. The good correspondence between the IL scores and behavioral scores (a correlation of 0.91), shown in Fig 2b, indicate the IL score is indeed a good measure for speech intelligibility assessment.

3.2. Transitory changes in signal intelligibility

The proposed model is a general intelligibility prediction system that can compute an ongoing estimate of signal intelligibility. Fig 3a illustrates the drop in predicted IL scores of a speech utterance as babble noise at 0dB is introduced at

[1] The phonological clustering of speech using this model shall be described further in a future publication

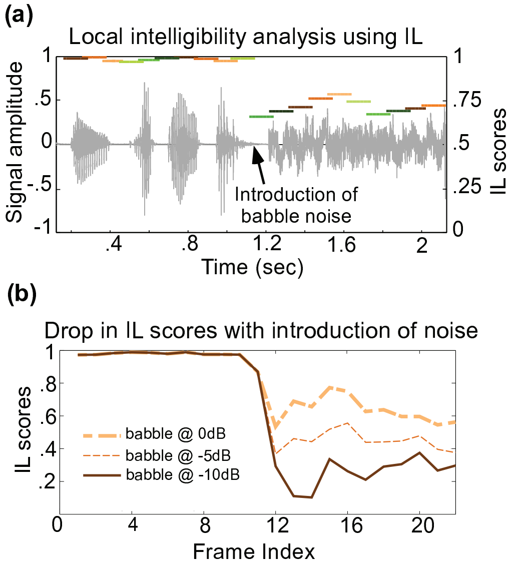

Fig. 3. Illustration of tracking of transitory changes in signal intelligibility

around 1.1sec. Further, as the noise level is increases, the drop in IL scores is higher as illustrated in Fig 3b (displayed as mean IL scores computed over 25 TIMIT sentences with babble introduced at 1.1sec at different SNR levels).

By performing a local-level signal analysis, the proposed model can generate IL scores at sub-lexical level (phonemic or sub-syllabic) hence tracking the transitory changes in intelligibility. To test the validity of this premise, noise is introduced at an arbitrary time instant to a speech utterance and the IL scores are used to estimate the noise introduction time. In a simple scheme, two contiguous IL scores that fall below a nominal threshold are taken as the indication of noise introduction time. The threshold is chosen as the average IL score for the given noise type and the SNR level, pre-computed on a different set of 25 TIMIT speech utterances. The correlation coefficient between actual and predicted time instants is 0.83 [2] confirming that the model is indeed able to predict transitory changes in intelligibility with reasonable accuracy.

3.3. Robustness of IL scores

An often overlooked issue in speech intelligibility models is robustness of the methods to a variety of speech corpora. Existing methods which do not make use of reference clean speech templates but rely on average profiles of speech or generic templates clearly fall short in the robustness aspect.

To validate the robustness of our approach, IL scores for a range of noise conditions (different noise types at different SNR levels) are compared across three speech corpora: TIMIT, NTIMIT and SWITCHBOARD. Since the statistical models are trained on TIMIT, the correlation of IL scores between TIMIT-NTIMIT and TIMIT-SWITCHBOARD is taken as the measure of robustness. The strong agreement between IL scores across the different databases, shown in Table 1 (average correlation of 0.98), suggests the proposed model generalizes well to multiple speech corpora.

Database	Noise Types					
	babble	factory	pink	street	train	white
NTIMIT	0.95	0.99	0.98	0.98	0.98	0.98
SBOARD	0.96	0.99	0.99	0.96	0.99	0.98

Table 1. Robustness of IL scores across different databases

4. CONCLUSIONS

We proposed a novel approach to speech intelligibility assessment that combines biologically-inspired acoustic analysis of peripheral and cortical processing, along with statistical modeling of the inherent variability present in speech. The approach has two unique advantages: (i) does not require reference clean speech templates (ii) enables local-level tracking of transitory changes in intelligibility. We showed that the proposed model predictions are robust across multiple speech corpora and strongly correlate with scores obtained from human listeners under a variety of realistic listening environments. Future work shall explore extensions of the model to incorporate contextual and syntactic linguistic information which can complement the current microscopic analysis with a more macroscopic analysis of speech intelligibility. The proposed model may be of interest also in applications such as speech enhancement where identifying regions distorted by noise (and the amount of of distortion) is crucial for the effectiveness of the algorithms.

5. REFERENCES

[1] 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.

[2] ANSI-S3.5-1969-R1978, “Methods for the calculation of the articulation index,” 1969.

[3] H. J. Steeneken and T. Houtgast, “A physical method for measuring speech-transmission quality,” *J Acoust Soc Am*, vol. 67, no. 1, pp. 318–26, 1980.

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

[5] 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.

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

[7] 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.

[8] S.C. Chen and D.H. Li, “Modified linear discriminant analysis,” *Pattern Recog*, vol. 38, pp. 441–443, 2005.

[2] A more sophisticated scheme might yield a better correlation