A Microscopic Approach to Speech Intelligibility Prediction using Auditory Models

Mahdie Karbasi, Dorothea Kolossa
Ruhr-Universität Bochum, Germany, Email: {mahdie.karbasi, dorothea.kolossa}@rub.de

Abstract

Speech intelligibility prediction has long been a challenging topic in the field of speech signal processing. Recently, the so-called microscopic intelligibility models have attracted a lot of attention, since they promise to be more precise in estimating intelligibility and in diagnosing problems due to specific phoneme confusions. These methods attempt to predict the listener's response to a speech signal on a word-by-word or phoneme-by-phoneme basis. For this purpose, they typically use a model which simulates the signal processing in the auditory periphery. In this paper, we develop a new speech intelligibility prediction approach which is based on HMMs of auditory features. The approach uses discriminance scores in these auditory HMMs as intelligibility features, and it is evaluated for speech in speech-shaped noise at different SNRs. Also we compare the results obtained by our model with one state-of-the-art macroscopic model.

Introduction

Speech intelligibility is a measure of the percentage of the words that can be recognized correctly by human listeners. To reduce the amount of necessary but time consuming and costly listening tests during the design and optimization of new speech processing systems, it would be important to be able to estimate this measure more reliably than it is currently possible.

Generally there are two different approaches for predicting the intelligibility of speech, the so-called *macroscopic* and *microscopic* approaches. Macroscopic models are the first and most widely-used group of methods, which employ macroscopic features of the signal like its long-term frequency spectrum. Some of the well-known early macroscopic measures including the articulation index (AI) [1] and the speech transmission index (STI) [2] were proposed to deal only with linear degradations like linear filtering or additive noise. Later, measures like the speech-based envelope power spectrum model (EPSM) [3] were designed to also cope with more complex distortions. The short time objective intelligibility (STOI) [4] measure and the mutual information k-nearest neighbor [5] approach are other recently proposed tools for estimating speech intelligibility.

In contrast to these macroscopic models, recently there have been many developments regarding more fine-grained microscopic models. In [6] the authors use a glimpsing model and an automatic speech recognition system (ASR) to investigate the speaker-related features affecting the intelligibility. In [7], the combination of an auditory model and the dynamic time warping algorithm is used to predict the accuracy of normal-hearing and hearing-impaired listeners in recognizing single words. Modeling the auditory nerves to simulate the neurograms and using image similarity metrics is a relatively different approach, which is used in [8] to predict the human phoneme discrimination scores. In this paper we propose a new ASR-driven measure in order to predict the intelligibility of speech more accurately. For this purpose, we briefly investigate the effectiveness of auditory models as a feature extraction stage for the ASR system in comparison to standard Mel frequency cepstral coefficients (MFCC). Based on this ASR setup the performance of our proposed measure in predicting the human recognition accuracy is then compared with that of the STOI [4] measure.

Microscopic Model

An automatic speech recognition system is the main part of our microscopic model. Therefore implementing an accurate ASR system plays a major role in predicting the speech intelligibility precisely. In this work the main consideration for implementing the feature extraction stage of the ASR system was having a model of the human auditory system. This factor will later enable us to investigate the effects of hearing loss on the intelligibility of speech by integrating the hearing impairment into the whole model. Therefore we chose auditory models as the core of the feature extraction stage. According to our own tests and to previous works [7], the model proposed by Dau [9] yields promising results within an ASR system. The outputs of the modulation filterbank in [9] are therefore used as features for an ASR system based on hidden Markov models (HMMs) in the following. For ASR, the JASPER system [10] has been used.

Discriminative Intelligibility Measure

For extracting the proposed discriminative intelligibility measure, an accurate ASR system needs to be trained at first. At test time, degraded speech is given to this system to obtain the recognized words as the ASR output.

In the next step, for each word in the sentence, the recognition output W_R and the ground-truth transcription W_T are used to compute a discriminance score. As this score, we propose to use an HMM-based log likelihood ratio (HLLR), given by

$$HLLR = \log \frac{P(S|W_R)}{P(S|W_T)}. \quad (1)$$

Here, S represents the degraded speech features, $P(S|W_R)$ is computed based on the HMM of the recognized word and $P(S|W_T)$ based on the HMM of the true word.

Finally, the HLLR is used as the feature for a support vector machine (SVM) classifier, which solves a two-class problem, classifying each word as recognized either correctly or incorrectly by human listeners.

Experimental Setup and Results

The Grid corpus [11] has been used in all following experiments. In Table 1 the performance of the ASR system with the aforementioned feature extraction is being compared in terms of speech recognition accuracy, training and testing speaker-independent HMMs on clean signals. As can be seen, the accuracy of auditory model (AM) features after applying principal component analysis (PCA) is high enough in comparison to MFCCs (which have always had high performance in speech recognition tasks). However, since MFCCs are not modeling the human auditory system explicitly, they cannot be easily used to integrate hearing impairment into our model. Therefore, these results let us conclude that auditory models are a reasonable choice for the feature extraction here.

Table 1: HMM recognition accuracy for different features and dimensionalities (D)

Num. of Gaussians	AM (D = 108)	AM+PCA (D = 39)	MFCCs (D = 39)
1	92.50	96.13	96.17
3	95.76	97.73	97.81
5	96.45	98.10	98.13

Finally, for assessing the proposed intelligibility estimation approach, we have trained condition-dependent HMMs on noisy versions of the Grid corpus, as described in [11], and we have used those to extract the HLLR scores. Also, the STOI was computed for these files. SVMs were trained to classify words as likely to be recognized correctly or incorrectly by human listeners. Separate SVMs are used for both of these intelligibility features, once for the STOI and once for the newly introduced discriminance measure, the HLLR. Table 2 shows the accuracy of these SVMs in predicting human performance on a word-by-word basis. As can be seen, the HLLR slightly outperforms the STOI measure, which has been proven to have accurate results in different SNRs and different noise conditions, in all tested conditions.

Table 2: SVM accuracy in predicting human listening results using two different objective measures

SNR (dB)	Measure	
	STOI	HLLR
-4	86.10	87.25
0	94.52	94.72
4	97.92	97.97
6	98.32	98.35
Clean	99.25	99.25
Mean	95.22	95.51

Conclusions

In this paper we have proposed a new intelligibility measure, the HMM-based log likelihood ratio (HLLR), which can be used to predict speech intelligibility of single words more accurately than the STOI measure without needing the clean version of the signal. It will also allow us to model hearing impairments physiologically, which will be a central topic of our future work.

Acknowledgments

This research has received funding from the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n°[317521]. The authors would like to thank Jon Barker for providing a noisy version of the Grid database with listening test results.

References

[1] French, N. and Steinberg, J.: Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19 (1947), 90–119

[2] Steeneken, H. J. M. and Houtgast, T.: A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am. 67 (1980), 318–326

[3] Jørgensen, S. and Dau, T.: Predicting speech intelligibility based on the envelope power signal-to-noise ratio after modulation-frequency selective processing. J. Acoust. Soc. Am. 130 (2011), 1475–1487

[4] Taal, C. H., Hendriks, R. C. and Heusdens, R.: An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. J. Acoust. Soc. Am. 130 (2011), 3013–3027

[5] Taghia, J. and Martin, R.: Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing. IEEE/ACM Trans. Audio, Speech and Lang. Process. 22 (2014), 6–16

[6] Barker, J. and Cooke., M.: Modelling speaker intelligibility in noise. Speech Communication 49.5 (2007), 402-417

[7] Jürgens, T., Fredelake, S., Meyer, R. M., Brand, T. and Kollmeier, B.: Challenging the speech intelligibility index: macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners. Interspeech (2010), 2478-2481

[8] Hines, A. and Harte, N.: Speech intelligibility prediction using a neurogram similarity index measure. Speech Communication 54.2 (2012), 306-320

[9] Dau, T., Kollmeier, B. and Kohlrausch, A.: Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102 (1997), 2892-2905

[10] Kolossa, D., Zeiler, S., Vorwerk, A. and Orglmeister, R.: Audiovisual speech recognition with missing or unreliable data. Proc. AVSP (2009)

[11] Cooke, M., Barker, J., Cunningham, S. and Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120 (2006), 2421-2424