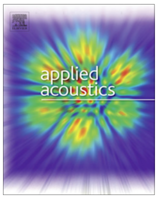
Applied Acoustics 129 (2018) 408–416

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust



Estimation of binaural speech intelligibility using machine learning

Kazuhiro Kondo [*], Kazuya Taira [1]

Graduate School of Science and Engineering, Yamagata University, 4-3-16 Jonan, Yonezawa, Yamagata 9928510, Japan

ARTICLE INFO

Article history:
Received 6 April 2017
Received in revised form 28 August 2017
Accepted 5 September 2017
Available online 9 September 2017

Keywords:
Speech intelligibility
Binaural speech
Objective estimation
Machine learning
Diagnostic rhyme test

ABSTRACT

We proposed and evaluated a speech intelligibility estimation method for binaural signals. The assumption here was that both the speech and competing noise are directional sources. In this case, when the speech and noise are located away from each other, the intelligibility generally improves since the auditory system can segregate these two streams. However, since intelligibility tests as well as its estimation is conducted based on monaurally-recorded signals, this potential increase in the intelligibility due to the segregation of sources is not accounted for, and the intelligibility is often under-estimated. Accordingly, in order to estimate the intelligibility taking into account this binaural advantage, we trained a mapping function between the subjective intelligibility and objective measures that account for the binaural advantage stated above. We attempted SNR calculation on (1) a simple binaural to monaural mixdown, which models the conventional estimation, (2) simple pooling of both binaural channels (pooled channel), (3) channel signal selection with the better SNR from left and right channels (better-ear), and (4) sub-band wise better-ear selection (band-wise better-ear). For the mapping function training, we tried neural networks (NN), support vector regression (SVR), and random forests (RF), and compared these to simple logistic regression (LR). We also investigated the sub-band configuration that gives the best estimation accuracy by balancing the frequency resolution and the amount of training data. It was found that the combination of the better-ear model and RF gave the best results, with root mean square error (RMSE) of about 0.11 and correlation of 0.92 in an open set test.

1. Introduction

Information communication using speech may potentially be conducted in all sorts of ambient noise conditions. For example, a lecture might be conducted either in a large classroom, or a small room with varying degree of reverberation. A conversation might be conducted with significant surrounding noise in a busy shopping mall. Accordingly, techniques for efficient and accurate speech communication quality assessment is necessary in order to conduct regular quality measurement to assure stable and sufficient speech communication over these various environments. Speech intelligibility is a measure that quantifies the accuracy of the perceived speech signals over a transmission channel, and thus is a crucial measure of the communication quality [1,2].

Speech intelligibility is measured using human subjects. The subjects listen to read speech samples, and identify the content of this speech. The content of the read speech may be syllables, words, or sentences. The subjects typically listen to each sample, and write down or select what they heard. The number of stimuli that needs to be evaluated needs to be large enough to cover all aspects of the language that is being tested, such as the phonetic context. The test also needs to include enough number of subjects so that the variation in the responses by subject are averaged out. Thus, intelligibility tests are generally time-consuming, and expensive.

Accordingly, numerous efforts to estimate the intelligibility without using human subjects have been conducted. One of the earliest examples of such estimation is the Articulation Index (AI) [3], which estimates the intelligibility from Signal-to-Noise Ratio (SNR) measurements within a number of frequency bands combined using a perceptual model. In a later effort, Steeneken and Houtgast proposed the Speech Transmission Index (STI) [4], which uses artificial speech signals communicated over the test channel to estimate the intelligibility of the received signal by measuring the weighted average modulation depth over frequency sub-bands.

However, most of these estimation methods estimate the monaural speech intelligibility using monaural signals. In the real world, however, the human subjects listen to speech signals using both ears, i.e., binaural signals. This can potentially improve the speech intelligibility since the human auditory system can

[*] Corresponding author.
E-mail address: kkondo@yz.yamagata-u.ac.jp (K. Kondo).
[1] Currently with Yamamoto Electric Corp.

http://dx.doi.org/10.1016/j.apacoust.2017.09.001
0003-682X/© 2017 Elsevier Ltd. All rights reserved.

potentially discriminate sources traveling from different directions. However, it is often the case that speech intelligibility estimation is conducted using monaural signals. This can lead to significant underestimation of speech intelligibility, especially when one can expect distinct noise sources to be located away from the target speech source.

Thus, there have been efforts to estimate the speech intelligibility from binaural signals. For example, Wijngarrden et al. attempted to improve the accuracy of STI on binaural signals [5]. They employed the inter-aural cross-correlogram, which is a plot of cross-correlation in each frequency band vs. the inter-aural delay, to adjust the contribution of each channel signal in each of the bands to the final Modulation Transfer Function (MTF), and showed that they can estimate the binaural intelligibility from binaural signals with their model at comparable accuracy that a conventional STI can predict the intelligibility on monaural signals.

We have also attempted to estimate the binaural speech intelligibility using binaural signals [6]. We calculated the frequency-weighted SNR for each channel, and applied the better-ear model [7] to this measure, and mapped this to intelligibility using a pre-trained logistic regression function. This seems to give a relatively accurate intelligibility estimation, with Root Mean Square Error (RMSE) and Pearson correlation of about 0.10 and 0.79, respectively. However, obviously there was still room for improvement.

In this paper, we attempted to make modifications to the better-ear model to improve the accuracy of these measures. We also introduce more sophisticated machine learning techniques to model the relation between subjective intelligibility and the objective measures [8]. As we will see, the selection of objective measure by sub-bands do not seem to be advantageous over selecting the channel signal as a whole. However, the use of random forests to map the objective measure to intelligibility significantly improves the accuracy of the estimated intelligibility, to a practical level.

This paper is organized as follows. In the next section, the binaural estimation method is outlined. This is followed by the estimation accuracy evaluation of the proposed method. Then, optimization of the filter bank used in the method is attempted and evaluated. Finally, conclusions and suggestions for further research is given.

2. Estimation of binaural speech intelligibility

Fig. 1 shows a block diagram of the proposed binaural speech intelligibility estimation method. In this method, we try to estimate the binaural intelligibility of a mixture of speech and noise source coming from various directions. We assume that not only the target speech, but also the noise is a directional source, such as a group of bystanders talking loudly from a specified direction, or an automobile or a train passing by from one direction to another.

We first train a mapping function between an objective measure calculated using the binaural signal to the subjective intelligibility. To train this mapping function, we compile a database of target speech traveling from various directions by convolving monaural target speech samples with the corresponding Head Related Transfer Functions (HRTFs). We also prepare noise sources from different directions by convolving this with the same HRTFs. Then, these two sources are mixed to compile a database of localized speech and noise with various azimuth combinations.

We conducted subjective intelligibility evaluations using the above database to compile a database of subjective intelligibility to use as supervisory signals in the training. The objective measure of each of the mixed signals is calculated, and the mapping function from this measure to the supervisory subjective intelligibility is trained. In our previous work [6], we used conventional logistic regression (LR) function for this mapping. However, we found that this function does not match the objective measure to the subjective intelligibility in the lower SNR range. Thus, in this paper, we attempted the use of some machine learning techniques, such as neural networks (NN), support vector regression (SVR), and random forests (RF) to improve the mapping accuracy at all SNR ranges.

The trained functions were then used to estimate the intelligibility of either a localized speech and noise combination used in the training (closed set testing), or speech mixed with noise not used during training (open set testing). Three different combinations of noise used for training and testing were attempted for the open set testing.

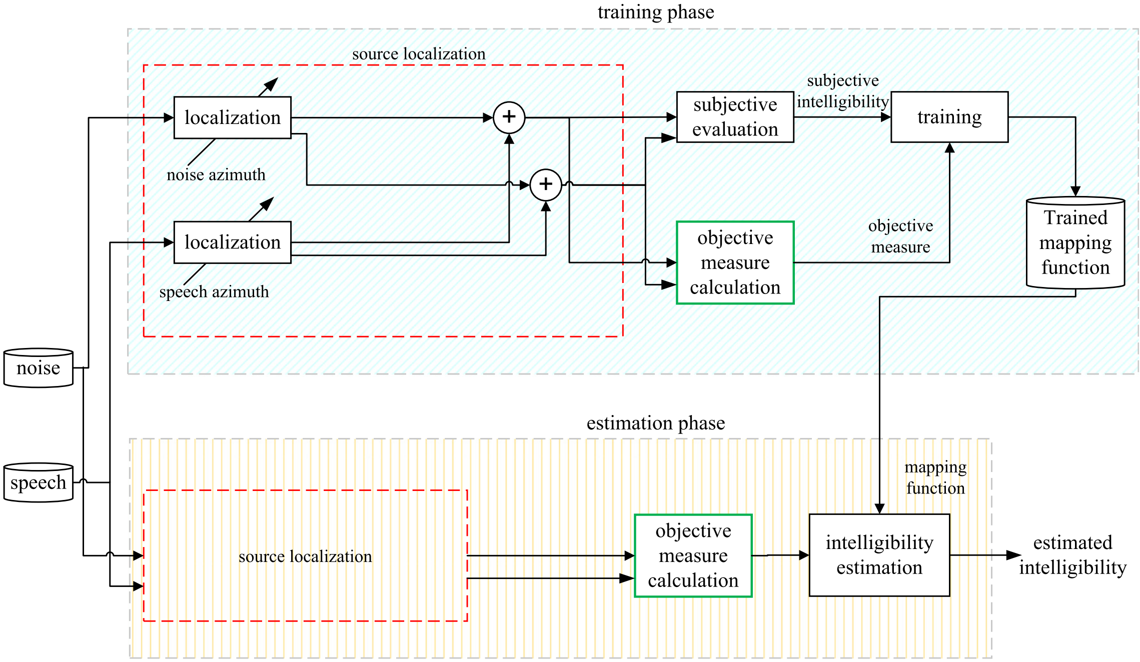


Fig. 1. Block diagram of speech intelligibility estimation.