Estimation of Binaural Intelligibility Using the Frequency-Weighted Segmental SNR of Stereo Channel Signals

Kazuya Taira* and Kazuhiro Kondo*
*Graduate School of Science and Engineering, Yamagata University, Yamagata, Japan
E-mail: {thh04707@st, kkondo@yz}.yamagata-u.ac.jp Tel: +81-0238-26-3312

***Abstract*— Most existing objective intelligibility prediction methods predict monaural intelligibility using monaural signals. These methods do not consider that a human can easily distinguish sounds arriving from different directions by sound heard in both ears. Therefore, intelligibility prediction using binaural signals that take this into account is necessary. Accordingly, speech samples with various source locations were prepared using binaural simulation. Subjective measurement tests were carried out on these samples. The frequency-weighted segmental SNR (fwSNRseg) was obtained using two different models for comparison: 1) a monaural model, which simply combines signals in both channels into one, and 2) the simple better ear model which uses the channel with the better SNR. Binaural intelligibility is estimated by applying regression analysis on the fwSNRseg and the subjective measurement results.**

Intelligibility was predicted by applying the resultant regression function on fwSNRseg of the test sample. We compared the estimation precision of the two models with binaural signals. The estimation precision of the better ear model yielded higher correlation with subjective scores than that of the monaural model by approximately 0.2 in a closed test as well as two open sets.

I. INTRODUCTION

Speech intelligibility is a measure of speech quality, and is used in many fields concerned with speech. Speech intelligibility can be measured using either a subjective or an objective evaluation method. Because subjective evaluation requires significant effort and cost, an objective evaluation method is desirable. However, most of existing objective intelligibility prediction methods predict monaural intelligibility using monaural signals. These methods do not consider that a human can easily distinguish sounds arriving from different directions by using sound heard in both ears. This can potentially improve the speech intelligibility. Therefore, it is necessary to predict the intelligibility using binaural signals in order to take this into account.

Wijngaarden et al. have attempted to improve the accuracy of the speech transmission index (STI) on binaural signals [1], which they termed the binaural STI. They employed the inter-aural cross-correlogram to adjust the contribution of each band to the final Modulation Transfer Function (MTF) estimate. They have shown that the binaural STI can improve the estimation accuracy of binaural signals.

In this paper, we propose an objective speech intelligibility measurement method which can improve the estimation accuracy of binaural signals by using the fwSNRseg, which has been shown to correlate highly with subjective intelligibility [2]. We initially employ a simple better ear model with the fwSNRseg, and find out how far this model can effectively improve the prediction accuracy on binaural signals.

II. THE PROPOSED OBJECTIVE INTELLIGIBILITY ESTIMATION PROCEDURE

The overall configuration of the proposed objective intelligibility estimation procedure is shown in Fig. 1. We selected to use the fwSNRseg as our objective measure since we have previously shown that this measure shows the highest correlation with subjective intelligibility under various conditions compared to other measures [2]. The signal used in [2] was monaural. Since we will be using stereo signals in this work, we have two objective measures for both of the channels, and so we simply chose to select the higher value out of the two, i.e., the better ear model. Then the selected fwSNRseg was mapped to the speech intelligibility using a pre-trained regression model.

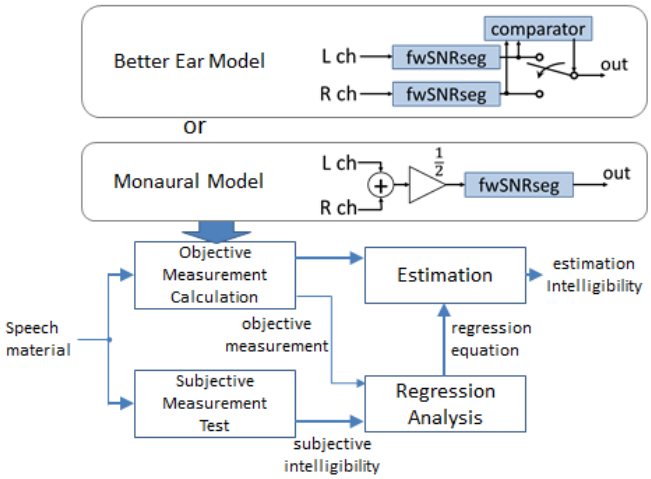

Fig. 1 Intelligibility estimation procedure

A. *Better Ear Model and Monaural Model*

Since the Human Auditory System (HAS) can discriminate sources when their locations are distant, the intelligibility of the speech source may improve if this source is located away

from other interfering sources. This needs to be taken into account when estimating intelligibility using objective measures. However, most objective intelligibility estimation does not take this into account, and tend to underestimate the intelligibility when speech is located away from interference. We will initially attempt to take these characteristics of the HAS into account using a very simple hearing model.

The better ear model simply selects the objective measure with higher value out of the two measures available for the two channels, i.e. left and right. Although this model is extremely simple, it was shown to give good match with subjective intelligibility. This may be because HAS is processing binaural signals in a similar manner, at least in the peripheral stages.

In order to compare the better ear model to the conventional monaural signal based estimation, we also attempted to estimate intelligibility on the objective measure calculated with the mixed-down single channel. We shall call this the monaural model. The mixed-down mono signal is simply the average of the two channels.

B. *The fwSNRseg measure*

We chose to use the fwSNRseg as the objective measure since this correlated well with subjective intelligibility. The fwSNRseg measure applies weights defined in the Articulation Index (AI) standard to each frequency band, which corresponds to the sensitivity of the HAS to the energy in each band [2,4]. The fwSNRseg values were computed using (1), where $W(j,m)$ is the weight placed on the j^{th} frequency band, K is the number of bands, M is the total number of frames in the signal, $X(j,m)$ is the critical-band magnitude (excitation spectrum) of the clean signal in the j^{th} frequency band at the m^{th} frame, and $X'(j,m)$ is the corresponding spectral magnitude of the degraded signal in the same band [2,4].

$$\text{fwSNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)\log_{10}\dfrac{X^2(j,m)}{\left(X(j,m)-X'(j,m)\right)^2}}{\sum_{j=1}^{K} W(j,m)}. \qquad (1)$$

C. *Regression analysis*

A regression model was trained using the objective measure output of the two models described above in the training data, and the corresponding subjective intelligibility as the supervisory signal. We chose to use the logistic regression and the polynomial regression (3rd order).

III. Subjective Intelligibility Test

Subjective intelligibility was measured on binaural signals in order to be used as the supervisory signal in the regression model training, and to evaluate the accuracy of the intelligibility during testing.

A. *The Japanese DRT*

Subjective speech intelligibility was measured using the Japanese Diagnostic Rhyme Test (JDRT) [3]. The JDRT is a forced two-to-one selection test, in which the subject is presented a word speech and given a choice of two words from which the subject must choose from. The word pairs are minimal pairs which only differ by the initial consonant. An example of a word pair is "**h**ashi" and "**k**ashi." Due to the simplicity of this test, it was shown that even an untrained subject can produce stable test results.

B. *Speech Sample*

We chose 30 word pairs, 60 words from the full JDRT word list. The word speech was spoken by one female. The following noise samples were added to this word speech. Babble noise from the Signal Processing Information Base (SPIB) database from Rice University [5]. We also selected fan coil noise and local train noise from the JEIDA noise database [6]. The level of the added noise was scaled to generate noise added samples with SNRs of 0, –6, and –12 dB, respectively. Both the noise and speech were localized by convolving with the Head Related Transfer Function (HRTF) of a KEMAR Mannequin, which is available from MIT Media Lab [7]. Speech and noise samples were localized at 0°, ±45°, and ±90° on the horizontal plane. We also included diotic samples as the control condition, in which the same level was played out from both channels, and no localization was applied. The total number of samples with different combinations of word speech, SNR, and localization was 16,200.

C. *Subjective measurement result*

The subjective evaluation was conducted using headphones to play out the test samples. Six subjects with normal hearing in their early twenties participated in the tests. Fig. 2 is an excerpt of the results. The noise used in the tests in Fig. 2 was babble noise, with SNR set at –12 dB. The noise was localized at –45°, 0°, and 45°, respectively. In all cases, the intelligibility is lowest when the noise and speech directions match, as expected.

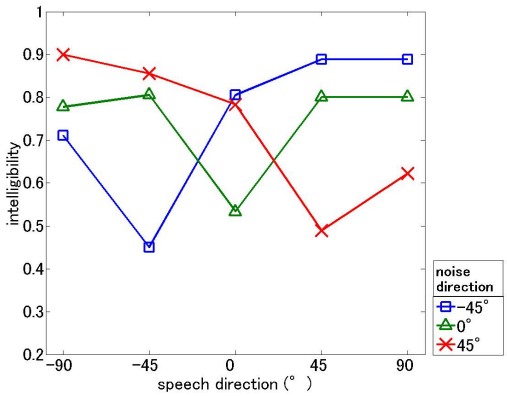

Fig. 2 subjective measurement result

IV. Intelligibility Estimation

Speech intelligibility was estimated by applying a regression model trained using the subjective intelligibility described in the previous chapter, and the fwSNRseg. We first attempted to estimate the intelligibility of the samples in the training set (closed set testing). Then, we estimated intelligibility of samples with noise not used for training (open set testing). Intelligibility was estimated using both the better ear model and the monaural model for comparison.

A. Closed set estimation result

A regression model was trained using fwSNRseg as the independent variable, and the binaural speech intelligibility as the dependent variable. Both logistic regression as well as a polynomial regression (3rd order) was used based on the results of a pretest. The logistic regression model was trained using the glm (generalized linear model) function, and the polynomial regression model was trained using the lm (linear model) function in the R programming environment. The distribution of fwSNRseg calculated with the better ear model and the monaural model vs. the subjective binaural intelligibility is shown in Figs. 3 and 4, respectively. The trained regression functions are also shown, where the solid lines show the logistic regression and the dashed lines show the polynomial regression. The correlation between the binaural subjective intelligibility and the fwSNRseg with the better ear model was 0.648, while it was 0.510 with the monaural model. Thus, the better ear model more effectively reflects the binaural subjective intelligibility.

The trained regression models were used to map the fwSNRseg to intelligibility. Figs. 5 and 6 show the subjective vs. estimated intelligibility using the monaural and the better ear model, respectively. Logistic regression was used in these cases. The better ear model seems to show more plots close to the diagonal line compared to the monaural model, which refers to more accurate estimations.

Table I shows the Root Mean Square Error (RMSE) and the Pearson correlation between the subjective and estimated intelligibility. The better ear model shows about 0.2 higher correlation and slightly lower RMSE compared to the monaural model. The better ear model gave more accurate estimation than the monaural model, both with logistic and polynomial regression. This most likely was because the fwSNRseg with the better ear model was able to reflect the subjective intelligibility especially at lower SNR. The monaural model fails to give lower fwSNRseg at lower SNR regions. The polynomial regression gives slightly lower RMSE (0.004) and slightly higher correlation (0.02) than logistic regression with the better ear model. However, this difference is not statistically significant.

TABLE I
ESTIMATION ACCURACY OF THE CLOSED TEST

Model	Regression	RMSE	Correlation
Monaural	Logistic	0.121	0.548
	Polynomial	0.121	0.549
Better Ear	Logistic	0.093	0.768
	Polynomial	0.089	0.788

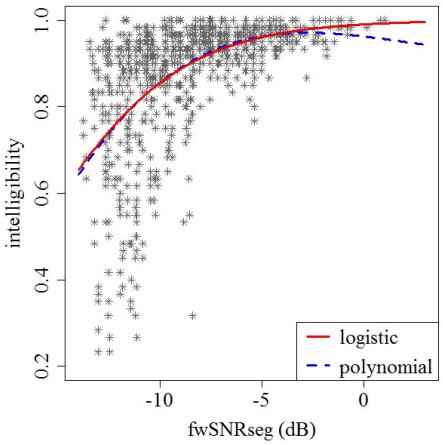

Fig. 3 fwSNRseg vs. subjective intelligibility scatter diagram and regression curve with the monaural model

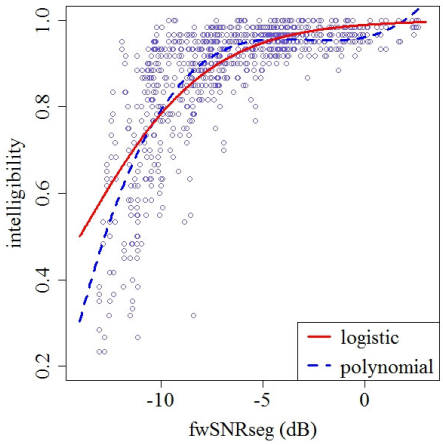

Fig. 4 fwSNRseg vs. subjective intelligibility scatter diagram and regression curve with the better ear mode

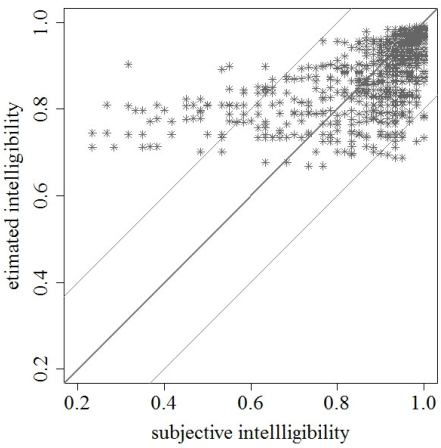

Fig. 5 subjective vs. objective intelligibility scatter diagram with the monaural model

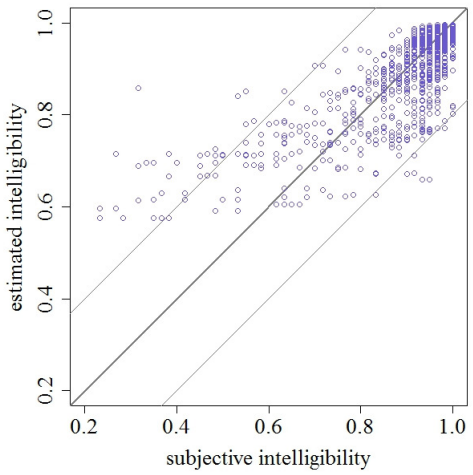

Fig. 6 subjective vs. objective intelligibility scatter diagram with the better ear model

B. *Closed set estimation result*

We also conducted open noise type tests using regression models trained on two types of noise, and testing on unknown noise. We tested three training schedules listed in Table II. The RMSE and the correlation between the subjective and estimated intelligibility is shown in Tables III and IV, respectively. As can be seen, the open test results are not significantly different from closed test results. The RMSE of the better ear model is about 0.03 smaller, and the correlation is higher by about 0.25 compared to the monaural model in most cases.

Logistic regression and polynomial regression seems to give similar accuracies in all cases except for the monaural model in schedule 3, in which the polynomial model gives significantly lower correlation. This seems to show that the monaural model may give unstable results depending on the conditions. In the conditions tested, we did not experience such outlier cases with the better ear model.

TABLE II
ESTIMATION ACCURACY OF THE CLOSED TEST

Noise Type	Schedule 1	Schedule 2	Schedule 3
Babble	Train	Test	Train
Fan		Train	Test
Rail	Test		Train

TABLE III
OPEN TEST RESULT (RMSE)

Model	Regression	RMSE		
		Schedule 1	Schedule 2	Schedule 3
Monaural	Logistic	0.150	0.125	0.083
	Polynomial	0.150	0.126	0.097
Better Ear	Logistic	0.119	0.099	0.066
	Polynomial	0.123	0.099	0.067

TABLE IV
OPEN TEST RESULT (CORRELATION COEFFICIENT)

Model	Regression	Correlation		
		Schedule 1	Schedule 2	Schedule 3
Monaural	Logistic	0.517	0.541	0.532
	Polynomial	0.516	0.539	0.392
Better Ear	Logistic	0.767	0.794	0.752
	Polynomial	0.794	0.800	0.721

V. CONCLUSIONS

We proposed and evaluated a binaural speech intelligibility estimation method using fwSNRseg and a simple binaural model, the better ear model. Regression models were trained using the fwSNRseg and the subjective intelligibility, and this regression was used to map the fwSNRseg to intelligibility. This simple model was shown to improve the estimation accuracy of binaural speech with interfering noise, both localized in separate directions compared to conventional monaural estimation. This accuracy improvement was also seen in cases where the noise used to train the regression models and noise used in the test data differ.

Since we used only three noise types in this initial experiment, we obviously need to test on more noise types. We also would like to use more sophisticated binaural hearing models than the simple better-ear model to improve the estimation. The Jeffress-Colburn model or the cocktail party processor [8] may be a good starting point. The addition of variables such as the Inter-aural Level Difference and Inter-aural Time Difference in the independent variable in the regression model might be another simple modification to our model which might improve the estimation accuracy.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Number 25330182.

REFERENCES

[1] S. J. Wijngaarden, R. Drullman, "Binaural intelligibility prediction based on the speech transmission index," *J. Acoust. Soc. Am.*, 123 (6), pp. 4514-4523, 2008.
[2] K. Kondo, "Estimation of Speech Intelligibility Using Objective Measures," *Elsevier Applied Acoustics*, 74 (1), pp. 63-70, 2013.
[3] K. Kondo, R. Izumi, M. Fujimori, R. Kaga and K. Nakagawa "Two-to-one selection-based Japanese speech intelligibility test," *J. Acoust. Soc. Japan*, 63 (4), pp. 196–205, 2007.
[4] J. Ma, Y. Hu and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, 125 (5), pp. 3387–3405, 2009.
[5] D. H. Johnson, P. N. Shami, "The signal processing information base," *IEEE Signal Processing Magazine*, pp. 36-43, 1993.
[6] S. Itahashi, "A noise database and Japanese common speech data corpus," *J. Acoust. Soc. Japan*, 47 (12), pp.951–953, 1991.
[7] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone." http://sound.media.mit.edu/resources/KEMAR/hrtfdoc.txt, 1994. MIT Media Lab Perceptual Computing - Technical Report #280.
[8] K. Iida, M. Morimoto, K. Fukudome, M. Miyoshi and T. Usagawa, *Spatial Hearing*, Tokyo: Corona, 2010.