A Speech Intelligibility Estimation Method Based on Hidden Markov Model

Deokgyu Yun, Jiseung Han, Jisu Son, Seung Ho Choi*
Dept. of Electronic and IT Media Engineering
Seoul National University of Science and Technology
Seoul, Korea
e-mail: shchoi@snut.ac.kr
**corresponding author*

Abstract—This paper proposes a speech intelligibility estimation method based on hidden Markov model (HMM) that is widely used for speech recognition. The HMM-based method is a kind of non-intrusive speech quality measurement, which means it operates without a reference speech signal. The log-likelihood score of HMM is converted to a normalized intelligibility score. We estimate the speech intelligibility of standard digital speech coders. The experimental results show that the proposed HMM-based method gives improved performance than the conventional non-intrusive speech intelligibility evaluation tool.

Keywords- *Speech intelligibility; hidden Markov model (HMM); digital speech coder; non-intrusive*

*****

I. Introduction

The need to improve speech intelligibility is increasing for a better communication. Accordingly, there are many research works for the speech intelligibility estimation [1], [2]. The evaluation method of speech intelligibility can be classified as intrusive or non-intrusive based on the need for a reference signal or not [2].

The standard methods of intrusive intelligibility evaluation are STOI (short time objective intelligibility) [3] and PESQ (perceptual evaluation of speech quality; ITU-T P.862) [4]. The STOI measures the correlation between the spectra of the reference and test speech. The PESQ is an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders. On the other hand, P.563 is widely used as a standard for non-intrusive speech quality assessment, which is based on models of the human vocal tract and the human perception of abnormalities in a voice signal [5].

In this paper, we propose a speech intelligibility estimation method based on hidden Markov model (HMM) that is widely used for speech recognition. The HMM-based method is a kind of non-intrusive speech quality measurement.

This paper is organized as follows: in Section 2, we describe the HMM-based intelligibility evaluation technique. In Section 3, we present evaluation results and compare with conventional methods. Conclusion is given in Section 4.

II. Speech Intelligibility Estimation

In this section, we briefly introduce the HMM theory and the proposed HMM-based intelligibility estimation method. Gaussian mixture model (GMM) can be thought of as a single state of HMM [6]. In other words, a state in HMM has a mixture of distributions, with the probability of belonging to a distribution represented by the observation probability. Each state in HMM have a set of emission probability defined as:

$$\mathrm{f}(X_1,\dots,X_m|GMM) = \prod_{i=1}^{M} f(X_i|GMM) \qquad (1)$$

where $\mathrm{f}(X_i|GMM)$ is a probability density function of an observation.

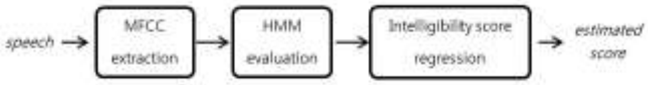

Figure 1. Example of a ONE-COLUMN figure caption.

HMM assumes that the hidden state follows the Markov property: conditional probability distribution of future states of the process depends only upon the present state, not on the sequence of events that preceded it. The outputs satisfy the Markov property too. The joint distribution of a sequence of states and observations is defined as:

$$\mathrm{P}(S_{1:T},X_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t) \quad (2)$$

where $X_{1:T}$, $P(S_1)$, $P(S_t|S_{t-1})$, and $P(X_t|S_t)$ represents observation sequences $\{X_1, X_2, \dots, X_T\}$, initial and transition probabilities, and output probability, respectively.

In order to evaluate the synthesized signal of the speech coder, the probability of occurrence is calculated. Accordingly, we can estimate speech intelligibility in non- intrusive configuration. The process of proposed HMM- based method is shown in Fig. 1, where MFCC means mel-frequency cepstral coefficient vector. The log-likelihood score of HMM is converted to an intelligibility score by linear regression analysis.

III. Experimental Results

We used TIMIT database for the evaluation [7]. The training and test data sizes are 4620 and 1680 utterances, respectively. The number of HMM states is 40 and the number of Gaussian mixtures is 64. As the input feature parameters, we used a 39-dimensional MFCC vector. The evaluation results are shown in Table 1.

In Table 1, DoD-CELP (Department of Defense code excited linear prediction) [8], MELP (mixed excitation linear prediction) [9] and LPC-10 (linear prediction coding) [10] are the standard low bit-rate speech coders. From this table, MELP shows the highest score in the HMM evaluation followed by DoD-CELP and LPC10. In STOI and PESQ evaluation, DoD-CELP shows the highest score followed by MELP and LPC10. However, in P.563 evaluation, MELP gets the highest score followed by LPC10 and DoD-CELP unlike the result of STOI and PESQ .

The correlation between the results of HMM and intrusive (STOI and PESQ) evaluation is higher than those of between the results of P.563 and intrusive evaluation. In conclusion, the result of HMM evaluation is similar to the result of conventional intrusive intelligibility evaluation. That means the proposed HMM-based method is more suitable for the speech intelligibility estimation than the conventional standard P.563.

TABLE I. COMPARISON OF HMM EVALUATION WITH CONVENTIONAL INTELLIGIBILITY EVALUATION TOOLS

Codec **Tool**	**DoD-CELP (4.8kbps)**	**MELP (2.4kbps)**	**LPC-10 (2.4kbps)**
STOI	0.90	0.88	0.83
PESQ	3.26	3.13	2.68
P.563	3.54	4.40	3.75
HMM	3.02	3.10	2.96

IV. CONCLUSION

In this paper, we described the speech intelligibility estimation method based on HMM. The method was applied to the intelligibility estimation of speech signals of low bit-rate speech coders. The correlation between the results of HMM and intrusive (STOI and PESQ) evaluation was higher than that obtained by the standard non-intrusive evaluation (P.563). We can conclude that the proposed HMM-based method is more suitable for the speech intelligibility estimation than the conventional standard method.

.

ACKNOWLEDGMENT

This study was supported by the Research Program funded by the SeoulTech(Seoul National University of Science and Technology).

REFERENCES

[1] Mowlaee, P., Saeidi, R., Christensen, M. G., and Martin, R., “Subjective and objective quality assessment of single-channel speech separation algorithms,” Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 00. 69-72, 2012

[2] Falk, T. H., Zheng, C., and Chan, W. Y., “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Transactions on Audio, Speech, and Language Processing, 18.7, pp. 1766-1774, 2010.

[3] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J., “A short-Time objective intelligibility measure for time-frequency weighted noisy speech,” Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, pp. 4214-4217, 2010.

[4] Rix, Antony W., et al., “Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs,” Acoustics, Speech, and signal Processing, 2001. Proceedings. (ICASSP’01). 2001 IEEE International Conference on. Vol. 2. IEEE, pp. 749-752, 2001.

[5] Malfait, L., Berger J., and Kastner M., “P.563—The ITU-T standard for single-ended speech quality assessment,” IEEE Transactions on Audio, Speech, and Language Processing 14.6, pp. 1924-1934, 2006.

[6] Rabiner, L. R., “A tutorial on hidden markov models and selected applications in speech recognition,” Proceedings of the IEEE, 77.2, pp. 257-286, 1989.

[7] Garofolo, John S., et al., “Timit acoustic-phonetic continuous speech corpus,” Linguistic Data Consortium, Philadelphia, 1993.

[8] Campbell, J. P., Welch, V. C., and Tremain, T. E., “The new 4800 bps voice coding standard,” Proc. Military and Government Speech Technology, pp. 735-737, 1989.

[9] Supplee, Lynn M., et al., “MELP: the new federal standard at 2400 bps,” Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. Vol. 2. IEEE, pp. 1591-1594, 1997.

[10] Tremain, T. E., “The government standard linear predictive coding algorithm: LPC-10,” *Speech Technology* 1.2, pp. 40-49, 1982.