Speech Intelligibility Enhancement Using a Statistical Model of Clean Speech

Petko N. Petkov[1], *W. Bastiaan Kleijn*[1,2], *Gustav Eje Henter*[1]

[1]Sound and Image Processing Lab, School of Electrical Engineering,
KTH-Royal Institute of Technology, Stockholm, Sweden
[2]School of Engineering and Computer Science, Victoria University of Wellington,
Wellington, New Zealand

Abstract

Subjective speech intelligibility deteriorates when the speech is presented in a noisy environment. A trivial solution to the problem is to increase the volume of the signal. In practice, however, this approach leads to fatigue and dissatisfaction on the side of the listener. It may also lead to excessive levels of output power that cause non-linear distortions in the audio equipment. Alternatively, speech intelligibility can be enhanced by modifying the speech under an energy-preservation constraint.

An effective and powerful paradigm is to select the modification by optimizing the output of an objective intelligibility measure. We consider the application of this paradigm to an intelligibility measure related to the classification error probability in an automatic speech recognition system (ASR). We target, therefore, primarily the application to recorded and synthetic speech and assume that a transcription of the message is available when modifying the speech signal. The approach is general and can be applied to a broad range of modification strategies. The high operating level of the proposed measure suggests that modification selection is less influenced by mismatches between subjective and objective intelligibility.

We consider two modification strategies (arguably orthogonal to each other) in combination with the proposed objective measure. The choice of these modifications is motivated by findings resulting from the analysis of human behavior [1] and prior work on the effect of cue enhancement on improving subjective intelligibility [2]. In particular, we consider i) long-term spectral modifications and ii) vowel to consonant energy ratio adjustment at the word level. Both strategies have been applied in recent speech pre-emphasis algorithms using lower-level intelligibility measures [3, 4].

The objective measure we adopt is of the form:

$$\boldsymbol{c}^\star = \underset{\boldsymbol{c}}{\operatorname{argmax}} \sum_{j=1}^{J} w_j \log\left(\mathrm{p}\left(\boldsymbol{f}_j \mid \boldsymbol{m}_j,\, \boldsymbol{c}\right)\right)$$
$$\text{s.t. } \boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{e}} = 1, \quad \boldsymbol{c} \geq 0, \tag{1}$$

where $\boldsymbol{f}_j,\ j \in \{1, \cdots, J\}$ are feature vectors extracted on a per-frame basis, $\boldsymbol{m}_j,\ j \in \{1, \cdots, J\}$ are Gaussian mixture models characterizing the features for particular phonetic units from the speech model in an ASR system pre-trained on clean speech, $\boldsymbol{c}$ represents the set of modification parameters, which can be viewed as gain factors in the spectral or the temporal domain depending on the modification, and w_j are weight factors that can be used, e.g., to manipulate the importance of the contribution of different phonetic groups. Energy preservation for the duration of the modification window is enforced by the equality constraint, which ensures that the sum of the normalized spectral-band or phone-unit energies $\bar{\boldsymbol{e}}^{\mathrm{T}} = [\bar{e}_1,\, \bar{e}_2,\, \cdots,\, \bar{e}_K]$ is preserved by the modification.

Subjective evaluation of the proposed approach was performed for an additive noise scenario (multi-speaker babble noise at $-3\mathrm{dB}$, SNR) using recorded speech and the spectral modification strategy at the word level. The results indicated significant and consistent improvement in subjective intelligibility for the modified over the original speech. Preliminary experiments with the temporal modification strategy revealed that it is possible to induce the desired behavior (relocating energy from vowels to consonants) and improve the intelligibility of the speech signal, as judged in informal subjective tests. Formal subjective evaluation of the temporal and the combination of the two strategies are planned to be performed shortly.

One specific challenge that needs to be overcome to ensure the robust performance of the proposed approach is related to the alignment between acoustic models and signal frames. While this information is *a priori* available for synthetic speech, it needs to be derived for recorded speech. We used forced alignment [5] between the transcription of an utterance and the clean speech waveform to achieve that. We observed that forced alignment often fails to locate correctly the individual phones and creates an error, which then propagates through the optimization process. A limitation of the approach was also established in regards to the spectral modification at the word level in relation to modifying fast speech. Edge effects together with sharp differences in the spectral gains between neighboring modification windows can effectively decrease the intelligibility of the modified below that of the original speech.

1. References

[1] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of Noise on Speech Production: Acoustic and Perceptual Analyses." *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, Sep 1988.

[2] V. Hazan and A. Simpson, "The Effect of Cue-Enhancement on Consonant Intelligibility in Noise: Speaker and Listener Effects," *Language and Speech*, vol. 43, pp. 273–294, 2000.

[3] B. Sauert and P. Vary, "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index and Audio Power Limitations," in *Proc. Europ. Sig. Proc. Conf.*, 2010, pp. 1919–1923.

[4] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A Speech Preprocessing Strategy for Intelligibility Improvement in Noise Based on a Perceptual Distortion Measure," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2012, pp. 4061–4064.

[5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009.