Enhancing Subjective Speech Intelligibility Using a Statistical Model of 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

The intelligibility of speech in adverse noise conditions can be increased by modifying the characteristics of the clean speech prior to its presentation. An effective and flexible paradigm is to select the modification by optimizing a measure of objective intelligibility. In this paper we apply this paradigm at the text level and optimize a measure related to the classification error probability. The proposed method was applied to a simple, yet powerful, band energy modification mechanism, under an energy preservation constraint. Subjective evaluation results provide a clear indication of a significant gain in subjective intelligibility. In contrast to existing methods, our approach is not restricted to a particular modification strategy and treats the notion of optimality at a level closer to that of subjective intelligbility.

***Index Terms*—speech modification, subjective intelligibility, statistical speech model**

I. Introduction

Speech signal modifications for improved subjective intelligibility represent an area of active research. Human strategies are analyzed to understand better the importance of changes in various speech descriptors during speech production in ambient noise environments [1], [2]. From the perspective of speech enhancement for engineering applications, a number of methods inspired by but not limited to human modification strategies have been proposed. These can be classified into two main groups according to the objective they aim to fulfill: i) rule-based methods with heuristic motivation, e.g., [3], [4], [5], [6] and ii) methods that optimize an objective measure, which correlates with subjective intelligibility, e.g., [7], [8]. In terms of consistency, robustness and performance evaluation, we perceive the second approach as more advantageous. The most fundamental intelligibility measure that one can use is the accuracy of the conveyed message. Existing measures aim to approximate this fundamental measure at lower levels of abstraction such as, e.g., short-term spectra.

In the context of i) speech synthesis, ii) playback of prerecorded media such as podcasts or iii) script-based presentations such as news readings and weather forecasts, it can be assumed that a word-level transcription of the message is available. More generally, in any situation where the clean speech signal can be accessed, a fairly accurate transcription can also be obtained with a state-of-the-art speech recognition system at the cost of increased computational complexity. Automatic speech recognition, however, is not the focus of this paper. In the following we assume that a transcription of the speech signal is available at the time of presentation.

When the speech waveform is *a-priori* available or the speech is synthesized, a large range of modification parameters become available. These include the expansion of the vowel space and the modification of phoneme time durations [1], [9], [10]. In addition, parameters operating at a higher level of abstraction, such as prosody modification and lexical changes are also accesible. The influence of such parameters will be reflected inadequately in the score of measures that operate at a low level of abstraction such as, e.g., the speech intelligibility index (SII) [11]. The aspiration for achieving optimality at a higher level and the possibility for applying various modifications within the same framework motivates us to explore from the start a more fundamental intelligibility measure. Using a model of clean speech from an automatic speech recognition (ASR) system, we formulate an objective measure as the likelihood of the noisy utterance, computed in terms of a sequence of feature vectors, conditional on the transcription and the speech model.

To place our work in perspective we first look at earlier methods. Most common are rule-based methods. In [3], [12] the energy of the speech signal is increased one spectral band after the other to achieve a target separation from the noise floor. Transients and consonants are emphasized in [4], [5] respectively. Intelligibility in the methods above is improved at the cost of increasing the total energy of the speech signal. For practical purposes, power limitations need to be applied to avoid hearing damage or distortions due to, e.g., non-linear effects in the audio equipment. Rule-based determination of the band-specific gains followed by energy preservation compensation is performed in [6].

More recently, the use of an objective intelligibility model (IM) for speech has been introduced [7], [8]. Speech IM-s [11], [13], [14] commonly operate at a relatively low level of abstraction. They extract features from the noisy and the clean speech signals and map them to an intelligibility score. The intelligibility measures considered in [7], [8] are based on the speech intelligibility index (SII) [11]. While SII has a number of limitations [15], it is attractive due to its simplicity. Being a function of the band-specific speech and noise power spectral levels, it facilitates the derivation of the optimal speech gain for each band given the power spectrum of the noise. It is also relatively straightforward to integrate a power constraint [7]. An alternative low-level measure is considered in [8] in addition to the SII. It is based on the front-end processing stage of a high-level IM [16], which in its entirety includes a missing-data speech recognizer.

While for most studies representative subjective validation of the proposed speech modification strategies is not available, the evaluation results indicate that the principle of using an intelligibility measure to select the optimal speech modification is well-motivated. To ensure high-level optimality and at the same time enable the use of a broad range of modification parameters, a more general measure of intelligibility is needed. We show that it is possible to define a practical measure that operates at the text level. We validate the proposed measure with a listening test using a band energy modification mechanism under an energy preservation constraint.

The remainder of this paper is organized as follows. Section 2 presents the philosophy behind the proposed modification framework. Practical considerations are discussed in Section 3 followed by experimental results in Section 4 and conclusions in Section 5.

II. A Paradigm for Increased Intelligibility

The objective of applying speech modifications prior to the presentation in a noisy environment is to enhance the capability of the speech signal to carry the message to the listener. Figure 1 presents a hierarchical view of the communication process. It indicates the levels of abstraction at which modifications can be applied to counteract the effect of distortions in the transmission channel. Starting from the top of the hierarchy, it is possible to adjust i) the choice of words used to represent the message, ii) the pronunciation of the selected words and iii) spectral properties unrelated to prosody, e.g., band-specific energy levels. Recent algorithms [4], [7], [8] perform modifications at the lowest level of abstraction.

To establish the benefit of a modification, ideally we would like to compare the intended and the perceived messages. If we now put this comparison in an optimization loop we can select the modifcation that maximizes the resemblance between the two messages. From a practical perspective, however, such a measure is not attractive due to the inherent delay and the need for a subject in the processing loop. The work-around in existing algorithms is to select the modification by optimizing the output of an objective speech inteligibility measure. The measures currently in use operate on the short-term spectra level of the listener side in the hierarchy representation from Figure 1. While this can be a computationally efficient strategy, it is tailored to a particular set of modification parameters and considers optimality at a low level of abstraction.

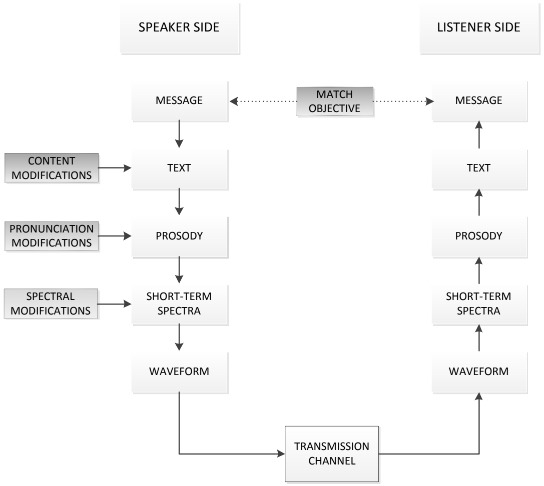

Figure 1. Hierarchical representation of speech communication.

For the purposes of this study we assume that a word-level transcription of the utterance to be presented is available. This allows us to perform matching at the text level, which is the highest level of abstraction (cf. Figure 1) at which at the current stage of technology an effective objective measure can be applied. The proximity to the message level suggests that modification selection based on optimization of the objective measure is less affected by mismatches between human intelligibility and the measure used. We focus explicitly on the scenario with modifying recorded speech as it facilitates the implementation and the validation of the proposed approach. While there are no explicit constraints on the type of distortion within the proposed framework, we consider the case of additive noise due to its broad practical implications.

The signal model for the additive noise scenario is given by

$$y\left(k\right)=x\left(k\right)+n\left(k\right), \tag{1}$$

where x is the speech, n is the noise, y is the additive mixture of the speech and noise source, and k indicates the time instant. In addition, let us introduce the operator Υ, which applied to a sequence of samples such as, e.g., $\mathbf{y}=\left[y\left(1\right)\,y\left(2\right)\,\cdots\,y\left(L\right)\right]$, produces a sequence of feature vectors, which we represent using matrix notation as $\mathbf{F}=\left[\mathbf{f}_1\,\mathbf{f}_2\,\cdots\,\mathbf{f}_J\right]^{\mathrm{T}}$, i.e.,

$$\mathbf{F}=\Upsilon\left\{\mathbf{y}\right\}. \tag{2}$$

The sample duration of $\mathbf{y}$ as well as the number and the dimension of the row vectors in $\mathbf{F}$ depend on the duration of the modifcation window and the choice of features. When using a speech model from an ASR system, most commonly the feature set is based on the mel frequency cepstral coefficients (MFCC) [17]. If we assume that the noise is wide sense stationary, the duration of the modifcation window represents a trade-off between the specificity of the modification and its flexibility. The longer the window gets, the less tailored to a particular sound or a short sequence of sounds the modifcation becomes. At the same time a longer window implies a broader range of modifcation possibilities to choose from. If the noise statistics are changing as well, the trade-off would include the validity of the estimated noise statistics for the length of the modifcation window.

We next introduce the objective function. From the perspective of the minimum classification error criterion our objective is to maximize

$$\mathrm{p}\left(\mathrm{t}|\mathbf{F},\,\mathbf{S},\,\mathbf{c}\right)=\frac{\mathrm{p}\left(\mathbf{F}|\mathrm{t},\,\mathbf{S},\,\mathbf{c}\right)\mathrm{p}\left(\mathrm{t},\,\mathbf{S},\,\mathbf{c}\right)}{\mathrm{p}\left(\mathbf{F},\,\mathbf{S},\,\mathbf{c}\right)},\tag{3}$$

where t is the transcription of the utterance, $\mathbf{S}$ is the speech model, which we take from an ASR system pre-trained on clean speech, and $\mathbf{c}$ contains the set of modification parameters. Alternatively, we can aim to minimize the probability of the set of all alternative transcriptions $\mathrm{t}_a^{(i)}$, $i\in\{1,\,2,\,\cdots,\,I\}$:

$$\sum_{i=1}^{I}\mathrm{p}\left(\mathrm{t}_a^{(i)}|\mathbf{F},\,\mathbf{S},\,\mathbf{c}\right)=\frac{\sum_{i=1}^{I}\left\{\mathrm{p}\left(\mathbf{F}|\mathrm{t}_a^{(i)},\,\mathbf{S},\,\mathbf{c}\right)\mathrm{p}\left(\mathrm{t}_a^{(i)},\,\mathbf{S},\,\mathbf{c}\right)\right\}}{\mathrm{p}\left(\mathbf{F},\,\mathbf{S},\,\mathbf{c}\right)},\tag{4}$$

whose cardinality I is a finite number. We combine the two criteria by adding the right-hand-side of (3) and the sign-inverted right-hand-side of (4). Sign inversion is necessary to express both formulations in terms of maximization. This allows us, after some equivalent transformations, to express the unconstrained objective function of the form:

$$\begin{aligned}\mathcal{O}\;&=\;\log\left\{\mathrm{p}\left(\mathbf{F}|\mathrm{t},\,\mathbf{S},\,\mathbf{c}\right)\right\}-\\&\quad\log\left\{\sum_{i=1}^{I}\left\{\mathrm{p}\left(\mathbf{F}|\mathrm{t}_a^{(i)},\,\mathbf{S},\,\mathbf{c}\right)\mathrm{p}\left(\mathrm{t}_a^{(i)}|\mathbf{S}\right)\right\}\right\},\end{aligned}\tag{5}$$

where the correct and the alternative transcriptions appear in separate terms. For practical purposes use of (5) is complicated by the need to maintain and evaluate the probabilities of the alternative transcriptions. It is possible to simplify the second term by including only the alternatives that achieve the highest scores. In the extreme case, we can omit all alternative transcriptions and focus only on the term containing the correct transcription. The optimization problem that we intend to solve is given, in its unconstrained form, by

$$\mathbf{c}^{*}=\mathrm{argmax}_{\mathbf{c}}\log\left\{\mathrm{p}\left(\mathbf{F}|\mathrm{t},\,\mathbf{S},\,\mathbf{c}\right)\right\}.\tag{6}$$

III. Practical considerations

The computation of $\mathrm{p}\left(\mathbf{F}|\mathrm{t},\,\mathbf{S},\,\mathbf{c}\right)$ from (6) requires access to the sequence of acoustic models associated with an utterance. While this information is *a-priori* available in a text-to-speech system, we need to derive it for a recorded utterance. We achieve this by performing forced alignment [17] between the transcription and the clean speech signal. The outcome of this operation provides us with an ordered list of acoustic models as well as segmentation boudaries and transition probabilities. Each feature vector $\mathbf{f}_j$ is, thus, associated with a particular state, which is represented by a Gaussian mixture model (GMM) from our speech model $\mathbf{S}$. Given the Markovian nature of $\mathbf{S}$ in ASR [17], $\mathrm{p}\left(\mathbf{F}|\mathrm{t},\,\mathbf{S},\,\mathbf{c}\right)$ is computed as

$$\mathrm{p}\left(\mathbf{F}|\mathrm{t},\,\mathbf{S},\,\mathbf{c}\right)=\prod_{j=1}^{J}\mathrm{p}\left(\mathbf{f}_j|\mathbf{m}_j,\,\mathbf{c}\right)\mathrm{p}\left(\mathbf{m}_j\rightarrow\mathbf{m}_{j+1}\right),\tag{7}$$

where $\mathbf{m}_j$ represents the state associated with frame j and $\mathrm{p}\left(\mathbf{m}_j\rightarrow\mathbf{m}_{j+1}\right)$ is the transition probability between the two states. Note that since we do not apply temporal modifcations, the product of the transition probabilities does not affect the optimization process.

We chose a low-level modification strategy to validate the proposed text-level intelligibility measure. It is based on the optimization of band energy gains, similar to [7], [8], under a total energy preservation constraint. We used a discrete Fourier transform filter-bank with a small number of channels (Cf. Table I). The bands are equally large on a mel-frequency scale to account for the spectral resolution of the human auditory system [18]. The set of modification parameters $\mathbf{c}^{\mathrm{T}}=\left[\mathrm{c}_1,\,\mathrm{c}_2,\,\cdots,\,\mathrm{c}_8\right]$ can now be used to express the energy preservation constraint. The optimization problem in its practical formulation becomes:

$$\begin{aligned}\mathbf{c}^{*}\;&=\;\mathrm{argmax}_{\mathbf{c}}\sum_{j=1}^{J}\log\left\{\mathrm{p}\left(\mathbf{f}_j|\mathbf{m}_j,\,\mathbf{c}\right)\right\}\\&\quad\text{s.t.}\quad\mathbf{c}^{\mathrm{T}}\bar{\mathbf{e}}=1,\;\;\mathbf{c}\geq0,\end{aligned}\tag{8}$$

where $\bar{\mathbf{e}}^{\mathrm{T}}=\left[\bar{\mathrm{e}}_1,\,\bar{\mathrm{e}}_2,\,\cdots,\,\bar{\mathrm{e}}_8\right]$ are the normalized energies of the signal (before modification) in the channels of the filter-bank for the duration of the modification window.

Table I
Top cut-off frequencies in the filter-bank.

f_c, [kHz]	0.26	0.61	1.10	1.77	2.68	3.93	5.65	8.00

The problem formulation from (8) can be solved with a standard package for constrained optimization. To speed up convergence, we used a finite difference approximation to the gradient of the objective function, which is not a closed-form function of $\mathbf{c}$.

The speech model was taken from an HTK-based ASR system [17] trained on 7138 utterances from the Wall Street Journal database. The signals were sampled at 16 kHz. We employed the CMU dictionary (version 0.6) [19]. The validation of the recognition system was conducted with utterances from the November 1992 CSR Speaker-Independent 5K Read Non-Verbalized Punctuation test set for which the recognizer achieved word correctness of 93.82 %.

The feature set per signal frame consisted of 12 Mel-frequency cepstral coefficients (MFCC) and the log energy, together with their first and second differentials. Cepstral mean normalization (CMN) [17] was applied to the final set of 39 features over the duration of the modification window. CMN mitigates the effect of the mismatch between the speech model $\mathbf{S}$ and the deviation from this model due to the modifications. It is, however, not critical for the operation of the algorithm.

IV. Experimental Results

We conducted a listening test with 30 utterances and eight subjects to validate the performance of the proposed approach. The clean speech recordings were taken from [20] and were spoken by a male native American English speaker. The speech material is composed of lists 44, 45 and 46 from the Harvard sentence database [21].

We mixed the speech signal with multi-speaker babble noise [22] at $-3\,\mathrm{dB}\,\mathrm{SNR}$. The noise level was chosen in a range where speech modifcations produced audible difference in the clarity of the enhanced speech under an energy preservation constraint. If the SNR is too low or too high it becomes difficult to measure the result of the modifcations with the limited number of subjects that we recruited.

Speech modifcation was performed at the word level, i.e., the length of the modification window adapted to the duration of each word. The algorithm had access to the mixture signal $\mathbf{y}$, which in practical terms means that we used an estimate of the noise power spectrum in each frame of the modification window. While the results can be viewed as an upper bound on the performance we expect to observe for on-line use, the oracle-like access to the noise statistics is not definitive for the performance of the algorithm. This is motivated with the global nature of the modification (it affects multiple frames), and the relatively stationary nature of the disturbance.

The test protocol can be summarized as follows. Fifteen of the utterances (after modification) and the remaining fifteen (before modification) were presented in noise to half of the subjects. The reverse combination was presented to the remaining subjects. Thus, no subject evaluated both the original and the modifed versions of the same utterance. Presentation within each of the two sets followed a randomized order where modified and unmodifed utterances alternated. After a presentation, each subject typed in what they heard.

To evaluate the performance we computed the recognition rate for a subject and utterance (modified or original) as the ratio of the correctly identified and the total number of words. We averaged these rates separately over the subjects rating the original and modifed versions. Averaging the utterance-specific mean recognition rates separately for all the modified and the original utterances produced the rates:

$$\bar{\mathrm{r}}_o = 0.379, \qquad \bar{\mathrm{r}}_m = 0.594$$

where the suffices o and m stand for original and modifed respectively. We applied the Wilcoxon signed rank test [23] to the series of per-utterance recognition rates corresponding to modifed and original utterances respectively. It revealed a significance at a level lower than 10^{-5}.

V. Conclusions

A general paradigm for enhancing the intelligibility of speech in noise, based on the optimization of an objective measure, was discussed. We formulated a fundamental and practical intelligibility measure as the likelihood of the noisy utterance given its transcription and a statistical model of clean speech. We applied the proposed approach to speech in multi-speaker babble noise using a simple but powerful band energy modification mechanism under an energy preservation constraint. The results from subjective evaluation confirmed the validity of the approach. A natural next step is to extend the set of modifications parameters and perform extensive validation for different noise types and SNR levels.

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] Y. Lu and M. Cooke, “Speech production modifications produced by competing talkers, babble, and stationary noise.” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, Nov 2008. [Online]. Available: http://dx.doi.org/10.1121/1.2990705

[3] J. W. Shin, W. Lim, J. Sung, and N. S. Kim, “Speech Reinforcement Based on Partial Specific Loudness,” in *Proc. Interspeech*, 2007, pp. 978–981.

[4] S. Yoo, J. R. Boston, J. D. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C.-C. Li, “Speech Enhancement Based on Transient Speech Information,” in *Proc. Appl. Sig. Proc. Audio and Acoust. Workshop*, 2005, pp. 62–65.

[5] P. S. Chanda and S. Park, “Speech Inteligibility Enhancement Using Tunable Equalization Filter,” in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2007, pp. 613–616.

[6] B. Sauert, G. Enzner, and P. Vary, “Near End Listening Enhancement with Strict Loudspeaker Output Power Constraint,” in *Intern. Workshop on Acoustic Echo and Noise Control*, 2006.

[7] 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.

[8] Y. Tang and M. Cooke, “Energy Reallocation Strategies for Speech Enhancement in Known Noise Conditions,” in *Proc. Interspeech*, 2010, pp. 1636–1639.

[9] J. C. Krause and L. D. Braida, “Acoustic Properties of Naturally Produced Clear Speech at Normal Speaking Rates.” *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 362–378, Jan 2004.

[10] B. Lindblom, A. Agwuele, H. M. Sussman, and E. E. Cortes, “The effect of emphatic stress on consonant vowel coarticulation.” *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3802–3813, Jun 2007. [Online]. Available: http://dx.doi.org/10.1121/1.2730622

[11] American National Standard, “Methods for the Calculation of the Speech Intelligibility Index,” 1997.

[12] B. Sauert and P. Vary, “Near End Listening Enhancement: Speech Intelligibility Improvement in Noisy Environments,” in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2006, pp. 493–496.

[13] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, “An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech,” *IEEE Trans. Audio, Speech, and Lang. Proc.*, no. 99, 2011, early Access.

[14] J. Ma and P. C. Loizou, “SNR Loss: A new objective measure for predicting speech intelligibility of noise-suppressed speech.” *Speech Communication*, vol. 53, no. 3, pp. 340–354, Mar 2011. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2010.10.005

[15] 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.*, vol. 125, no. 5, pp. 3387–3405, May 2009. [Online]. Available: http://dx.doi.org/10.1121/1.3097493

[16] M. Cooke, “A Glimpsing Model of Speech Perception in Noise.” *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, Mar 2006.

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

[18] B. C. Moore, *An Introduction to the Psychology of Hearing*. Elsevier Academic Press, 2004.

[19] C. M. University, “The CMU Pronouncing Dictionary,” ftp://ftp.cs.cmu.edu/project/speech/dict/.

[20] ITU-T Rec. P.Sup23, “ITU-T Coded-Speech database,” 1998.

[21] “IEEE Recommnded Practice for Speech Quality Measurements,” *IEEE Trans. Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[22] A. P. Varga, J. M. Steenneken, M. Tolimson, and D. Jones, “The noisex-92 study on the effect of additive noise on automatic speech recognition,” DRA Speech Research Unit, Tech. Rep., 1992.

[23] D. F. Bauer, “Constructing Confidence Sets Using Rank Statistics,” *J. Am. Stat. Assoc.*, vol. 67, pp. 687–690, 1972.