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 improved 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. Here we apply this paradigm at the text level and optimize a measure related to the classification error probability in an automatic speech recognition system. The proposed method was applied to a simple but 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, the proposed approach is not restricted to a particular modification strategy and treats the notion of optimality at a level closer to that of subjective intelligibility. The computational complexity of the method is sufficiently low to enable its use in on-line applications.

Index Terms: speech modification, subjective intelligibility, statistical model of speech

1. Introduction

Speech signal modifications for improved subjective intelligibility represent an area of active research. Human strategies, such as the Lombard effect, are analyzed to understand better the importance of changes in various speech descriptors during speech production in noisy environments [1, 2]. A number of methods inspired by but not limited to human modification strategies have been proposed for speech enhancement for engineering applications. These can be classified into two main groups: 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]. We favor the second approach as it provides a figure of merit indicative of the performance of the algorithm and facilitates the analysis of its behavior. The most fundamental intelligibility measure one can apply is the accuracy of the conveyed message. Existing measures commonly approximate this 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 audio books and podcasts, or iii) script-based presentations such as news broadcasts 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 typically 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], as well as gain adjustments in the spectral and the time domains at the phone, word or utterance levels. The influence of a large number of these modifications 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, conditioned on the correct transcription and the speech model.

To place our work in perspective we first look at earlier methods. Most common is the application of rule-based modifications. 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., nonlinear effects in the audio equipment. Rule-based computation of the band-specific gains followed by energy preservation compensation is performed in [6].

More recently, the use of objective speech intelligibility models (IMs) has been introduced [7, 8]. Speech IMs [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 levels, it facilitates the computation of the optimal speech gain for each band given the power 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],

The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 256230.

which in its entirety includes a missing-data speech recognizer.

The evaluation results for the above-listed modification algorithms indicate that 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. Section 3 describes the practical considerations related to the implementation of the proposed method. Section 4 presents the experimental results, followed by conclusions in Section 5.

2. A Paradigm for Increased Intelligibility

The objective with modifying speech prior to its 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, we would ideally like to compare the intended and the perceived messages. If we perform this comparison in an optimization loop in which we modify the speech, we can select the modification that maximizes the resemblance between the two messages. From a practical viewpoint, however, such a setup is not attractive due to the need for a subject in the processing loop and the inherent delay. The work-around is to select the modification by optimizing the output of an objective speech intelligibility measure. The measures currently in use operate predominantly on the listener-side short-term spectra level in the hierarchy of Figure 1. While this can be a computationally efficient strategy, it is tailored to a particular modification and considers optimality at a low level of abstraction.

In this study we assume that a word-level transcription of the presented utterance is available. This allows us to perform matching at the text level, which is the highest level of abstraction (cf. Figure 1) for 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 this objective measure is less affected by mismatches between subjective and objective intelligibility. 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 focus on additive noise due to its broad practical relevance.

The signal model for the additive noise scenario is given by

$$y_k = x_k + n_k, \tag{1}$$

where x is the speech, n is the noise, y is the additive mixture of the speech and noise sources, and k indicates the time

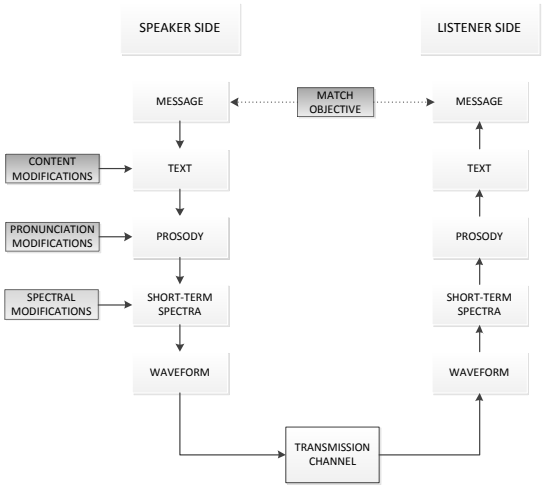

Figure 1: Hierarchical representation of speech communication.

instant. In addition, we introduce the operator Υ, which applied to a sequence of samples $\boldsymbol{y} = [y_1\, y_2\, \cdots\, y_L]$, produces a sequence of feature vectors, represented in matrix notation as $\boldsymbol{F} = [\boldsymbol{f}_1\, \boldsymbol{f}_2\, \cdots\, \boldsymbol{f}_J]^{\mathrm{T}}$, i.e.,

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

The number of samples in $\boldsymbol{y}$ as well as the number and the dimensionality of the row vectors in $\boldsymbol{F}$ depend on the duration of the modification window and the choice of features. When using a speech model from an ASR system, the feature set is most commonly based on Mel-frequency cepstral coefficients (MFCCs) [17]. If we assume that the noise is wide-sense stationary, the duration of the modification 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 the modification becomes. At the same time a longer window implies a broader range of possibilities for, e.g., energy redistribution. If the noise statistics are changing as well, the trade-off must include the accuracy of the predicted noise statistics for the length of the modification window.

We next introduce the objective function. From the perspective of minimizing the classification error, our objective is to maximize the posterior probability:

$$\mathrm{p}\left(t \mid \boldsymbol{F}, \mathcal{S}, \boldsymbol{c}\right) = \frac{\mathrm{p}\left(\boldsymbol{F} \mid t, \mathcal{S}, \boldsymbol{c}\right)\mathrm{p}\left(t \mid \mathcal{S}, \boldsymbol{c}\right)}{\mathrm{p}\left(\boldsymbol{F} \mid \mathcal{S}, \boldsymbol{c}\right)}, \tag{3}$$

where t is the correct transcription of the utterance, $\mathcal{S}$ is the speech model taken from an ASR system pre-trained on clean speech, and $\boldsymbol{c}$ contains the set of modification parameters. Alternatively and equivalently, from a theoretical viewpoint, we can minimize the probability of the set of all alternative transcriptions t_{a}^i, $i \in \{1, 2, \cdots, I\}$:

$$\sum_{i=1}^{I} \mathrm{p}\left(t_{\mathrm{a}}^i \mid \boldsymbol{F}, \mathcal{S}, \boldsymbol{c}\right) = \frac{\sum_{i=1}^{I}\left\{\mathrm{p}\left(\boldsymbol{F} \mid t_{\mathrm{a}}^i, \mathcal{S}, \boldsymbol{c}\right)\mathrm{p}\left(t_{\mathrm{a}}^i \mid \mathcal{S}, \boldsymbol{c}\right)\right\}}{\mathrm{p}\left(\boldsymbol{F} \mid \mathcal{S}, \boldsymbol{c}\right)}, \tag{4}$$

whose cardinality I is a finite number. The two criteria can be combined to formulate another theoretically equivalent optimization problem. This is achieved by taking the logarithm of the right-hand-side of (3) and adding it to the sign-inverted

logarithm of the right-hand-side of (4). The sign inversion is necessary to express both formulations as maximizations. After some manipulation of the resulting expression, the objective function is obtained of the form:

$$\mathcal{O} = \log\left\{\mathrm{p}\left(\boldsymbol{F} \mid t,\, \mathcal{S},\, \boldsymbol{c}\right)\right\} - \log\left\{\sum_{i=1}^{I}\left\{\mathrm{p}\left(\boldsymbol{F} \mid t_{\mathrm{a}}^{i},\, \mathcal{S},\, \boldsymbol{c}\right)\mathrm{p}\left(t_{\mathrm{a}}^{i} \mid \mathcal{S}\right)\right\}\right\}, \quad (5)$$

where the correct and the alternative transcriptions appear in separate terms. For practical purposes, the use of (5) is complicated by the need to maintain and evaluate the probabilities of all alternative transcriptions. It is feasible to approximate the second term by including only the alternatives that achieve the highest scores. In the extreme case, we can omit all alternative transcriptions. This is the scenario we consider here. The optimization problem we intend to solve is:

$$\boldsymbol{c}^{\star} = \underset{\boldsymbol{c}}{\operatorname{argmax}} \log\left\{\mathrm{p}\left(\boldsymbol{F} \mid t,\, \mathcal{S},\, \boldsymbol{c}\right)\right\}. \quad (6)$$

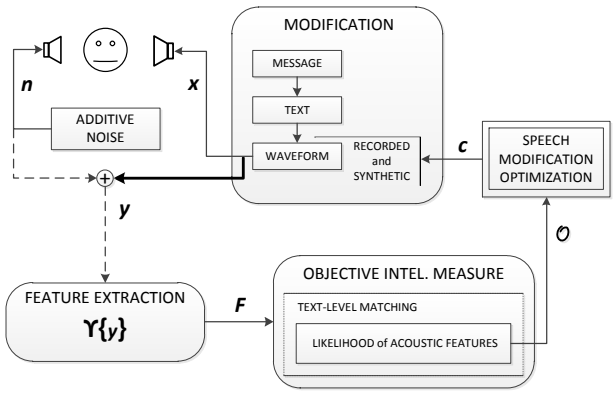

Figure 2: A diagram of the proposed system operation.

An illustration of the proposed approach is presented in Figure 2. A dashed line is used for the noise source to indicate that the noise waveform is not available and we work with an estimate of the statistics of the disturbance. A thick line is used for signal $\boldsymbol{x}$ to indicate that this is a closed-loop system in which only the optimally-modified signal is presented to the listener.

3. Practical considerations

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

$$\mathrm{p}\left(\boldsymbol{F} \mid t,\, \mathcal{S},\, \boldsymbol{c}\right) = \prod_{j=1}^{J}\mathrm{p}\left(\boldsymbol{f}_j \mid \boldsymbol{m}_j,\, \boldsymbol{c}\right)\mathrm{p}\left(\boldsymbol{m}_{j+1} \mid \boldsymbol{m}_j\right), \quad (7)$$

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

We chose to validate the proposed text-level intelligibility measure using a low-level modification strategy. It performs the optimization of band-energy gains, similar to [7, 8], under an energy preservation constraint. We used a discrete Fourier transform (DFT) filter-bank with a small number of channels (cf. Table 1). 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 $\boldsymbol{c}^{\mathrm{T}} = [c_1,\, c_2,\, \cdots,\, c_8]$ can now be used to express the energy preservation constraint. The optimization problem in its practical formulation becomes:

$$\boldsymbol{c}^{\star} = \underset{\boldsymbol{c}}{\operatorname{argmax}} \sum_{j=1}^{J}\log\left\{\mathrm{p}\left(\boldsymbol{f}_j \mid \boldsymbol{m}_j,\, \boldsymbol{c}\right)\right\} \quad \text{s.t. } \boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{e}} = 1, \quad \boldsymbol{c} \geq 0, \quad (8)$$

where $\bar{\boldsymbol{e}}^{\mathrm{T}} = [\bar{e}_1,\, \bar{e}_2,\, \cdots,\, \bar{e}_8]$ are the normalized band-energies in the original speech signal for the present modification window.

Table 1: Top cut-off frequencies ($\mathrm{f_c}$) in the filter-bank in kHz.

$\mathrm{f_c}$	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 standard software packages for constrained optimization. To speed up convergence, we used a finite difference approximation for the gradient of the objective function [19], due to its complicated dependence on $\boldsymbol{c}$. The small number of parameters and the *a-priori* established link between frames and GMMs ensure low computational complexity.

The speech model $\mathcal{S}$ 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) [20]. The recognition system was validated on 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 % in the absence of noise.

The feature set for a signal frame consisted of 12 MFCCs and the log energy, together with their first and second differentials, i.e., 39 features in total. The frame length was 25 ms and the frame update rate was 10 ms. Cepstral mean normalization (CMN) [17] was applied to the cepstral coefficients for the duration of the modification window. CMN compensates partially the deviation of the modified speech from the speech model $\mathcal{S}$, which was constructed for natural speech. This allows the system to perform more extreme spectral modifications.

4. Experimental Results

We conducted a listening test with 30 utterances and eight subjects to assess the performance of the proposed approach. The clean speech recordings were taken from [21] 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 [22]. We mixed the speech signal with multi-speaker babble noise [23] at -3 dB SNR. The noise level was

chosen such that subjective intelligibility is severely degraded but spectral energy redistribution is capable of producing an effective improvement.

Speech modification 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 $\boldsymbol{y}$, which means that we used an estimate of the true noise power spectrum in each frame of the modification window. In a practical application, the noise power spectra in the future cannot be estimated and must be predicted. While our results can be seen as an upper bound on the expected performance in on-line applications, oracle access to the noise spectra is not anticipated to be critical for multi-frame modifications and relatively stationary noise backgrounds, as considered here.

The subjective evaluation protocol can be summarized as follows. Fifteen of the utterances (modified) and the remaining fifteen (original) were presented in noise to half of the subjects. The reverse combination of modified and original material was presented to the remaining subjects. Thus, no subject evaluated both the original and the modified versions of the same utterance. Presentation within each of the two sets followed a randomized order where modified and original utterances alternated. After a presentation, each subject typed in the perceived message using blank spaces when unable to identify words.

To evaluate the method performance we computed the recognition rate for a subject and utterance as the ratio of the correctly identified to the total number of words. We averaged these recognition rates over the subjects and the utterances, separately for the original and modified versions, producing the mean recognition rates :

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

where the suffixes o and m stand for original and modified, respectively. We also applied the Wilcoxon signed rank test [24] to the series of per-utterance recognition rates corresponding to modified and original utterances respectively. It revealed a significance of the difference at a level lower than 10^{-5}.

5. 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 a 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 effective band-energy modification mechanism under an energy preservation constraint. The results from a subjective evaluation confirmed the validity of the approach. A natural next step is to extend the set of modifications parameters and perform experimental validation for different noise types and SNR levels.

6. 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. Woodland, *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] R. J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations*. Society for Industrial and Applied Mathematics (SIAM), 2007.

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

[21] ITU-T Rec. P.Sup23, "ITU-T Coded-Speech database," 1998.

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

[23] 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.

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