Rephrasing-Based Speech Intelligibility Enhancement

Mengqiu Zhang[1], *Petko N. Petkov*[1], *W. Bastiaan Kleijn*[1,2]

[1]School of Electrical Engineering, KTH-Royal Institute of Technology, Stockholm, Sweden
[2]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
mengqiu@kth.se, petkov@kth.se, bastiaan.kleijn@ecs.vuw.ac.nz

Abstract

Existing algorithms for improving speech intelligibility in a noisy environment generally focus on modifying the acoustic features of live, recorded or synthesized speech while preserving the phonetic composition (the message). In this paper, we present an algorithm for text-to-speech systems that operates at a higher level of abstraction, the message-level. We use a paraphrasing system to adjust the linguistic content of the intended message such that the speech intelligibility improves under noisy conditions. To distinguish the intelligibility among paraphrases, we use the numerical integration of a normalized log-likelihood function over different signal-to-noise conditions. Objective evaluation results show that the developed measure is able to distinguish the intelligibility among paraphrases. Results from subjective evaluation confirm the effectiveness of our objective measure.

Index Terms: speech intelligibility, message-level objective measure

1. Introduction

Recorded or synthetic speech from a text-to-speech system is increasingly used to deliver information in public announcement systems and private speaking systems embedded in the computer, mobile, and GPS (Global Positioning System) navigation devices. Unfortunately, the output speech is commonly presented in a noisy environment, and thus no guarantee can be made that the conveyed message is understandable. As in the listed applications the background noise signal cannot be reduced, a reasonable speech enhancement approach is to modify the original speech signal such that the speech intelligibility, a measure of the degree to which speech can be recognized, improves under adverse conditions.

Existing speech modification algorithms that aim to enhance speech intelligibility in a noisy environment using live, recorded, or synthetic speech can be categorized into two classes: i) rule-based and ii) objective-intelligibility-measure-based. The rule-based algorithms [1, 2] are normally inspired by the fact that human beings tend to change their vocal production to adapt to the ambient noise; the resulting speech is known as Lombard speech [3, 4, 5]. In contrast, objective-intelligibility-measure-based algorithms [6, 7, 8, 9, 10] modify the speech signal with the aim to increase the value of an objective measure. Examples of this measure are the speech intelligibility index (SII) [11], the glimpse proportion (GP) [12], the log-likelihood measure [6], or the perceptual distortion (PD) [7]. The above two classes of algorithms both involve the intentional modification of acoustic properties presented in speech segments (e.g., amplitude envelope, temporal fine structure, fundamental frequency, shifts of formants, and so on). Therefore, the improvement of speech intelligibility generally comes at the expense of increasing the signal distortion. Such distortion often causes discomfort and fatigue to the listener, particularly in presence of severe ambient noise. This suggests that alternative approaches should be considered.

We draw our inspiration from the approach humans take: in noisy environments humans rephrase and repeat the message. Hence, instead of modifying the acoustic features of the speech signal, we propose a message-level (MSG-level) speech intelligibility enhancement algorithm, where the intelligibility is improved by adjusting the linguistic content of the intended message. It capitalizes on the linguistic properties that may be informative for understanding the intended message under the noise condition [13, 14]. Two operations are required in the proposed MSG-level enhancement algorithm: 1) paraphrasing, and 2) distinguishing intelligibility.

The paraphrasing step is to provide the most likely set of alternatives for a target word or phrase in a given context. One approach is to apply ranking methods [15, 16, 17] on a list of candidates extracted from knowledge sources, such as WordNet [18] and Roget [19]. Another approach is based on automatically extracted paraphrase dictionary using bilingual parallel corpora [20]. Since the translation between bilingual parallel corpora preserve the meaning of the original message while may use different words to convey the meaning, the acquired paraphrase dictionary is suitable for us to generate paraphrases.

A primary objective of this paper is to distinguish the intelligibility among paraphrases given a particular noise condition. To this end, we start by deriving a MSG-level objective measure assuming that transcriptions of the paraphrased speech signals are available. The objective measure is based on the paradigm we have used for phoneme-level speech intelligibility enhancement [6], where the improvement of intelligibility is achieved by maximizing the likelihood of desired speech features. One consideration in the present work is that the transcripts are different among the paraphrases and thus, the corresponding speech models are different. Therefore, the resulting likelihoods are not directly comparable. Our analysis shows that the paraphrases exhibit different behaviors in terms of log-likelihood of the speech features over a range of signal-to-noise ratios (SNRs). We exploit these differences in our MSG-level objective measure to distinguish the most intelligible phrase in noisy environment. In an experiment where we select the best of three paraphrases, our objective measure matches the subjective results for about 80% of the data.

2. MSG-Level Speech Intelligibility Enhancement

Figure 1 presents the signal model of MSG-level speech intelligibility enhancement for text-to-speech system when the recorded announcements are rendered in a noisy environment.

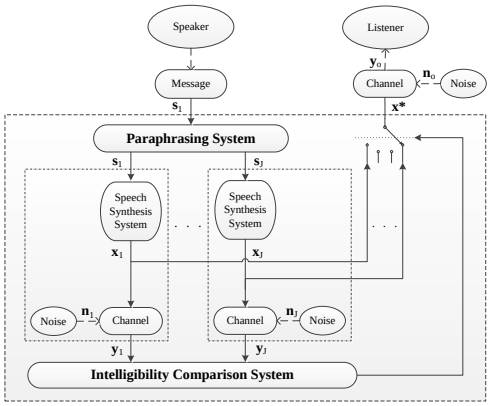


Figure 1: *Signal model for MSG-level speech intelligibility enhancement.* $\mathbf{s}_1$ *is the original message.* $\{\mathbf{s}_j\}_{j,\in[1,\,J]}$ *is a set of paraphrases including the original one.* $\{\mathbf{x}_j\}_{j\in[1,\,J]}$ *is a set of speech signals corresponding to the paraphrases.* $\{\mathbf{n}_j\}_{j\in[1,\,J]}$ *is a set of noise signals.* $\{\mathbf{y}_j\}_{j\in[1,\,J]}$ *is a set of noise-contaminated speech signals.* $\mathbf{x}^*$ *is the identified most intelligible paraphrase under a certain noisy condition.*

We assume that the word-level transcription of the intended message is available in this work. This transcription is denoted as $\mathbf{s}_1$ in Figure 1. First the paraphrasing system generates possible paraphrases $\{\mathbf{s}_j\}_{j\in[2,\,J]}$ to express the intended message in alternative ways, where j is the index of the paraphrase. Then a set of speech signals are obtained from the speech synthesizer, indicated by $\{\mathbf{x}_j\}_{j\in[1,\,J]}$. Presenting the clean speech signals in a noisy environment (here we consider the additive noise scenario), we have the additive mixtures of the speech and the noise signal, $\{\mathbf{y}_j\}_{j\in[1,\,J]}$,

$$\mathbf{y}_j = \mathbf{x}_j + \mathbf{n}_j. \tag{1}$$

To effectively convey the intended message in a noisy environment, a comparison system is needed to distinguish the intelligibility among paraphrases and manipulate the paraphrasing system to output this most intelligible expression, indicated by $\mathbf{x}^*$ in Figure 1. In the paper we use an existing paraphrasing system to generate alternative phrases while focusing on the development of the objective measure to evaluate intelligibility.

2.1. Paraphrasing

Phrase-based approaches using parallel bilingual corpora have been shown to produce high quality results for paraphrase extraction [20, 21, 22], word sense disambiguation [23], and statistical machine translation [24, 25]. In paraphrasing systems, pairs of translated sentences from a bilingual corpus are aligned, and the English phrases that share a common foreign language phrase as a translation are considered to be potential paraphrases. We adopt the strategy described in [20], which further applies syntactic constraints to the phrase extraction heuristics, to generate paraphrases as inputs to the intelligibility comparison system.

2.2. Distinguishing Intelligibility

A MSG-level objective measure to distinguish the intelligibility among paraphrases is established in this section. The algorithm proposed here assumes that the statistics of the noise signal is known (or can be estimated).

Let $\mathbf{F}$ represent some parametric description of the speech signal. Given the phonetic transcription of the speech signal of each paraphrase, denoted by $\mathbf{t}_j$, we write the probability distribution of the features as $\mathrm{p}\left(\mathbf{F}_{\mathbf{x}_j}|\mathbf{t}_j\right)$. Let $\mathbf{F}_{\mathbf{n}}$ represent some parametric features of the known environmental noise. Then the feature distribution of the noise-contaminated speech $\mathbf{y}_j$ can be represented by $\mathrm{p}\left(\mathbf{F}_{\mathbf{y}_j}|\mathbf{F}_{\mathbf{x}_j},\mathbf{F}_{\mathbf{n}_j}\right)$. Given $\mathbf{F}_{\mathbf{y}_j}$ and a corresponding speech model ν_j for decoding the speech features, the probability over the space of the decoded transcription $\boldsymbol{\tau}_j$ is $\mathrm{p}\left(\boldsymbol{\tau}_j|\mathbf{F}_{\mathbf{y}_j},\nu_j\right)$. The probabilities of all possible transcriptions must add up to one:

$$\mathrm{p}\left(\mathbf{t}_j|\mathbf{F}_{\mathbf{y}_j},\nu_j\right) + \sum_{\boldsymbol{\tau}_j,\boldsymbol{\tau}_j\neq t_j} \mathrm{p}\left(\boldsymbol{\tau}_j|\mathbf{F}_{\mathbf{y}_j},\nu_j\right) = 1. \tag{2}$$

The higher value of the first term on the left side of the equation, the smaller value of the second term, which implies a higher degree of intelligibility. Therefore, we use the ratio of the two terms as an indicator of intelligibility. Applying Bayes' rule to the probabilities, the ratio reads

$$\mathcal{R}_j = \frac{\mathrm{p}\left(\mathbf{F}_{\mathbf{y}_j}|\mathbf{t}_j,\nu_j\right)\mathrm{p}\left(\mathbf{t}_j|\nu_j\right)}{\sum_{\boldsymbol{\tau}_j,\boldsymbol{\tau}_j\neq t_j}\mathrm{p}\left(\mathbf{F}_{\mathbf{y}_j}|\boldsymbol{\tau}_j,\nu_j\right)\mathrm{p}\left(\boldsymbol{\tau}_j|\nu_j\right)}. \tag{3}$$

Notice that there is a monotonic relation between $\mathcal{R}_j$ and its numerator $\mathrm{p}\left(\mathbf{F}_{\mathbf{y}_j}|\mathbf{t}_j,\nu_j\right)\mathrm{p}\left(\mathbf{t}_j|\nu_j\right)$. As the denominator in (3) involves evaluating the likelihoods given all alternative transcriptions, which dramatically increases the computation cost, for practical purposes, it is feasible to focus only on the term containing the correct transcription while omitting all alternative transcriptions. Then, we take logarithm of the numerator and use it as an objective measure of intelligibility,

$$\mathcal{L}_j = \log\left(\mathrm{p}\left(\mathbf{F}_{\mathbf{y}_j}|\mathbf{t}_j,\nu_j\right)\right) + \log\left(\mathrm{p}\left(\mathbf{t}_j|\nu_j\right)\right). \tag{4}$$

However, the transcripts are different among phrases and the corresponding speech models, $\{\mathcal{L}_j\}_{j\in[1,\,J]}$ are not directly comparable.

We propose a relative method where the degradation of a speech in a noisy environment is considered as a key element. For each phrase $\mathbf{x}_j$, varying the SNR results in a sequence of $\{\mathcal{L}_j^i\}_{i\in[1,\,I]}$, where i is the index of SNR condition with $i = 1$ referring the noise signal and $i = I$ the clean speech. Since the range of log-likelihood among phrases varies widely, we normalize it by rescaling its range to $[0,\,1]$, where the minimum of 0 corresponds to the likelihood of the noise signal and the maximum of 1 to the likelihood of the clean signal [26].

As the intelligibility of a noise-contaminated speech signal will not be worse than a noise signal, we normalize the likelihood such that its minimum is given by the noise signal. On the other hand, when a small amount of noise is added to the speech, $\mathcal{L}_j^i$ might greater than that of the clean speech ($\mathcal{L}_j^I$) as the noise smooths the spectrum of the speech, especially in speech-shaped noise. However, perceptually, the noisy speech would not be more intelligible than clean speech. Therefore, we set the maximum of $\{\mathcal{L}_j^i\}_{i\in[1,\,I]}$ to one. Then, the normalized likelihood $\widetilde{\mathcal{L}}_j^i$ is

$$\widetilde{\mathcal{L}}_j^i = \min\{\frac{\mathcal{L}_j^i - \mathcal{L}_j^1}{\mathcal{L}_j^I - \mathcal{L}_j^1},\,1\}, \quad i\in[1,\,I],\,j\in[1,\,J]. \tag{5}$$

As the second term in (4) is not affected by the noise, it is canceled out in this normalization. Thus, $\mathcal{L}_j^i$ can be simplified to calculating the log-likelihood of the correct transcription after observing the $\mathbf{F}_{\mathbf{y}_j^i}$,

$$\mathcal{L}_j^i = \log\left(\mathrm{p}\left(\mathbf{F}_{\mathbf{y}_j^i}|\mathbf{t}_j,\nu_j\right)\right). \tag{6}$$

We can see that $\{\widetilde{\mathcal{L}}_j^i\}_{i\in[1,\,I]}$ is actually a function of SNR with its value within $[0,\,1]$. Let $\widetilde{\mathcal{L}}_j(s),\ s\in[\mathrm{SNR}_1,\,\mathrm{SNR}_I]$ denote the function. For any given SNR s_0, the closer the $\widetilde{\mathcal{L}}_j(s_0)$ is to 1, the more robust the speech is to the noise. To obtain a single measure of robustness for each phrase, we integrate this measure of robustness over all SNR. That is, we define the integral $\int_s \widetilde{\mathcal{L}}_j(s)ds$ as the MSG-level objective measure to indicate the robustness of the phrase to the noisy condition. The numerical approximation of the integration is then given by

$$\mathcal{O}_j = \sum_{i=1}^{I} \widetilde{\mathcal{L}}_j^i \Delta s_i. \tag{7}$$

Hence, under a certain noisy environment, the most intelligible phrase $\mathbf{x}_{j^*}$ (indicated by $\mathbf{x}^*$ in Figure 1) in the set $\{\mathbf{x}_j\}_{j\in[1,\,J]}$ is identified, where

$$j^* = \arg\max_j \mathcal{O}_j,\ j^*\in[1,\,J]. \tag{8}$$

2.3. Hidden Markov Model Based Implementation

In this subsection, we discuss the evaluation of (7) in a practical system. In particular we consider the specification of the speech model ν and the type of the speech feature $\mathbf{F}$.

The speech model we use builds upon hidden Markov models (HMMs) [27]. Gaussian mixture models (GMMs) are employed to approximate the distributions of speech features associated with the states of the HMMs. As the pronunciation of a word or subword unit, such as a phone, depends heavily on the context, we use the context-dependent speech model consisting of three emitting states.

The Mel-frequency cepstral coefficients (MFCCs) [28] are commonly used as features in ASR system. In such applications, the MFCC feature set usually consists of 12 static MFCCs and 24 dynamic MFCCs, the first and second-order differentials of the static MFCCs, with cepstral mean normalization (CMN) [28].

The dynamic MFCCs and CMN reduce the effect of noise and compensate partially for the difference between the noise-contaminated speech and the clean speech. This is essential in an ASR system. However, the main purpose of this work is to reveal sensitivity of the likelihood to the noise. Thus, the effect of noise should not be cancelled out. Therefore, we use 12 static MFCCs only and they are not normalized by CMN.

After defining ν and $\mathbf{F}$, we can calculate $\widetilde{\mathcal{L}}_j^i$ in (7). As the frame-number over a phone and the number of phones over a phrase both vary, we apply the invariance correlation, *i.e.*,

$$\mathcal{L}_j^i \approx \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L_k}\frac{1}{L_k}\log\Big(\mathrm{p}\Big(\mathbf{F}_{\mathbf{y}_j^i}^{k;l}|t_j^k,\nu_j\Big)\Big), \tag{9}$$

where k is the phoneme index, l is the frame index within the k-th phone, t_j^k is the k-th phoneme in the transcription $\mathbf{t}_j$, and $\mathbf{F}_{\mathbf{y}_j^i}^{k;l}$ are the features of l-th frame within the k-th phone of the noisy speech.

3. Experimental Results

In this section, we describe the objective and subjective evaluation setup and present their evaluation results.

3.1. Objective Evaluation Setup

To test the proposed system, we conducted objective evaluations under three noise conditions, *viz.*, airport, car, and speech-shaped, where the noise signals are from Aurora 2.0 database. We designed 15 English phrases, five for each noise condition. Based on the 15 phrases, a paraphrase dictionary was generated using the technique described in [20]. As the generated dictionary still contains many paraphrases that are either simply repetitions or inappropriate in the given context, we manually checked the grammaticality and meaning preserving nature of a paraphrase. The resulting speech material consists of 43 phrases in total, five sets of phrases for each type of noise with two or three paraphrases in each set.

The clean speech signals were synthesized with a well-known commercial text-to-speech system, using a female speaker at 16kHz sampling rate. We first normalized both the speech signal and the noise signal such that their L^2 norms are 1, *i.e.*, $\|\mathbf{x}_j\| = 1$ and $\|\mathbf{n}_j\| = 1$. Then we mixed them for a sequence of SNR values varying from 100 dB to -100 dB. In order to obtain the desired SNR, we used the FaNT tool [29] with the "-u -m snr 8khz" option. The noise realization is different for each phrase while it remains the same when varying the SNR.

For each set, the most intelligible phrase was then distinguished using (8) where the log-likelihood is calculated from (9) given the static MFCCs extracted from the mixed signal and the context-dependent phoneme models.

The phoneme model is pre-trained from an HTK-based automatic speech recognition (ASR) system [29] on clean speech. The training data consists of 7138 utterances from the Wall Street Journal (WSJ0) database [30], at a sampling frequency rate of 16 kHz. We use the CMU dictionary (ver. 0.6) [31] for forced-alignment between phonetic transcription and waveforms. The analysis frame length is 25 ms and the updated frame length is 10 ms.

3.2. Subjective Evaluation Setup

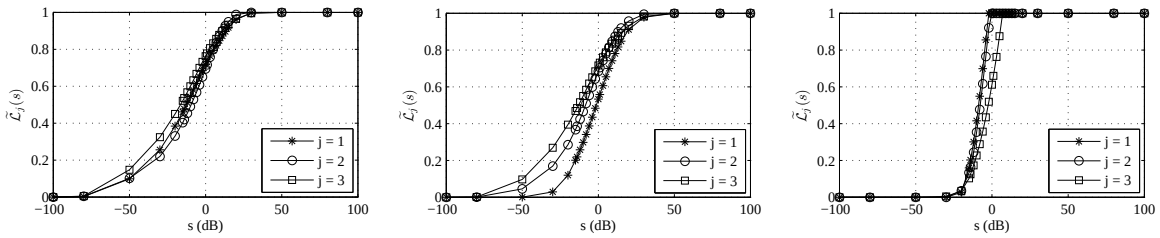
We conducted subjective experiments to evaluate the performance of the objective measure. The listening experiment was carried out by six non-native English speakers aged between 25 and 32. We used the same speech material as that in the objective evaluation, 43 phrases in three noise conditions. The SNR level for the listening test was fixed at -4 dB for the short phrases with no more than five words and -2 dB for the long phrases.

The phrases categorized in the same set have the same meaning and share some common words. If they are played in order, the subjects may benefit from previous experience of listening. Therefore, we scrambled the 43 utterances and presented them in a way that the consecutive utterances belonging to the same set are not played in order. After playing out an utterance, the subjects were asked to type in the perceived message.

We compute the recognition rate for an utterance as an averaged ratio of the correctly identified (misspelt words are included, such as 'regieon' is considered as 'region') and the total number of words over the subjects. The higher the recognition rate the more intelligible the prase.

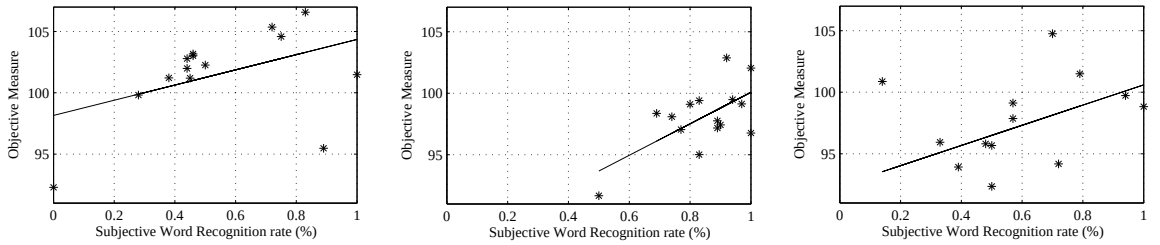
3.3. Evaluation Results

Figure 2 shows the objective evaluation results, the normalized log-likelihood of three sets of paraphrases [20] in three noisy conditions with SNR varying from -100 dB to 100 dB. The likelihood behaviors are similar in the three noisy conditions. With



(a) Airport Noise (b) Car Noise (c) Speech-Shaped Noise

Figure 2: *The normalized log-likelihood of features of the paraphrases [20] in three noisy conditions with SNR varying from -100dB to 100dB. (a) Airport Noise:* $j = 1$ *the school for ladies;* $j = 2$ *the school for women;* $j = 3$ *the school for girls. (b) Car Noise:* $j = 1$ *the very beginning;* $j = 2$ *the starting point;* $j = 3$ *the very first step. (c) Speech-Shaped Noise:* $j = 1$ *the tough problem;* $j = 2$ *the difficult problem;* $j = 3$ *the tricky problem.*



(a) Airport Noise (b) Car Noise (c) Speech-Shaped Noise

Figure 3: *Scatter plot between objective measure calculated by* (7) *and the averaged subjective word recognition rate for each phrase.*

the increase of SNR, the likelihood gradually increases from zero given by noise to one given by the clean speech. We can clearly see the difference of the area under each curve, which is utilized to distinguish the intelligibility among phrases. A more robust phrase gives a higher value of normalized likelihood under the same SNR condition and, thus, creates a larger area.

The objective measures of the phrases shown in Figure 2 and their subjective evaluation results are summarized in Table 1. We can see that a higher value of objective measure usually corresponds to a higher word recognition rate. In the whole experiment where we select the best of three paraphrases, our objective judgment matches the subjective results for about 80% of the data (12 matches out of 15 sets). A one-tailed binomial test shows that the match between objective judgment and the subjective judgment is statistically significant ($p < 0.001$).

The scatter plots in Figure 3 also show a positive correlation between the objective measure and the subjective word recognition rate. The solid lines are the linear regression lines that fit the data in a least squares sense. The corresponding correlation coefficients are 0.44, 0.64, and 0.48 for airport, car, and speech-shaped noise, respectively.

As the phrases are not professional designed, they might not be equally understandable to non-native subjects. This might induce bias to the subjective evaluation and weaken the correlation.

Table 1: *Objective Judgment vs. Subjective Judgment.*

Noise Type	j	Objective Measure $\mathcal{O}_j$	Word Recognition Rate (%)
Airport	1	103.2	0.46
	2	102.3	0.50
	3	106.6	0.83
Car	1	91.7	0.50
	2	99.5	0.94
	3	102.9	0.92
Speech	1	99.7	0.94
	2	98.8	1.00
Shaped	3	94.2	0.72

4. Conclusions

In this paper, a message-level paradigm was studied in an attempt to enhance the speech intelligibility without distorting the speech signal. We formulated the MSG-level objective measure as the numerical integration of the normalized log-likelihoods function obtained at different signal-to-noise conditions ranging from clean speech to noise. The proposed measure was tested under three noise conditions and a positive correlation between the objective measure and the subjective word recognition rate is observed. A natural next step is to improve the objective measure by taking into account the lexical effects and perform experimental validation on native English speakers for more noise types.

5. Acknowledgements

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.

We would like to express our gratitude to A. Prof. Chris Callison-Burch of the Computer Science Department at Johns Hopkins University for providing the paraphrase dictionary.

6. References

[1] E. Godoy and Y. Stylianou, "Unsupervised acoustic analyses of normal and Lombard speech, with spectral envelope transformation to improve intelligibility," in *Proc. of Interspeech 2012*, Portland Oregon, USA, September 2012.

[2] D. Huang and E. P. Ong, "Lombard speech model for automatic enhancement of speech intelligibility over telephone channel," in *Proc. of Int. Conf. on Audio Language and Image Processing (ICALIP)*, 2010, pp. 429–434.

[3] W. Van 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, September 1988.

[4] J. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, January 1993.

[5] 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, November 2008.

[6] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1035–1045, May 2013.

[7] 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. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, March 2012, pp. 4061 – 4064.

[8] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the glimpse propotion measure for improving the intelligibility of hmm-generated synthetic speech in noise," in *Proc. of Interspeech 2012*, Portland Oregon, USA, September 2012.

[9] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. of Interspeech 2010*, Makuhari, Japan, 2010, pp. 1636–1639.

[10] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, August 2010, pp. 1919–1923.

[11] *American National Standard: Methods for the Calculation of the Speech Intelligibility Index*, American National Std., ANSI S 3.5-1997, 1997.

[12] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, March 2006.

[13] R. Patel and K. W. Schell, "The influence of linguistic content on the Lombard effect," *J. Speech, Language and Hearing Research*, vol. 51, pp. 209–220, February 2008.

[14] D. Fogerty and L. E. Humes, "The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. 1490–1501, February 2012.

[15] D. McCarthy and R. Navigli, "SemEval-2007 task 10: English lexical substitution task," in *Proc. of the 4th Int. Workshop on Semantic Evaluations (SemEval-2007)*, 2007, pp. 48–53.

[16] S. Hassan, A. Csomai, C. Banea, R. Sinha, and R. Mihalcea, "UNT: SubFinder: Combining knowledge sources for automatic lexical substitution," in *Proc. of the Fourth Int. Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic, 2007, pp. 410–413.

[17] R. Sinha, "UNT-SIMPRANK: Systems for lexical simplification ranking," in *Proc. of First Joint Conf. on Lexical and Computational Semantics*, Montreal, Canada, June 2012, pp. 493–496.

[18] R. Poli, M. Healy, and A. Kameas, Eds., *Theory and Applications of Ontology: Computer Applications*. Springer, 2010, chapter 10 WordNet.

[19] M. Jarmasz and S. Szpakowicz, "Roget's thesaurus: A lexical resource to treasure," *arXiv preprint arXiv:1204.0258*, 2012.

[20] C. Callison-Burch, "Syntactic constraints on paraphrases extracted from parallel corpora," in *Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, October 2008, pp. 196–205. [Online]. Available: http://www.aclweb.org/anthology/D08-1021

[21] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Computational linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[22] R. Barzilay and K. R. McKeown, "Extracting paraphrases from a parallel corpus," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 50–57.

[23] W. A. Gale, K. W. Church, and D. Yarowsky, "Using bilingual materials to develop word sense disambiguation methods," in *Proc. of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992, pp. 101–112.

[24] C. Callison-Burch, P. Koehn, and M. Osborne, "Improved statistical machine translation using paraphrases," in *Proc. of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006, pp. 17–24.

[25] R. Zens and H. Ney, "Improvements in phrase-based statistical machine translation," in *Proc. of HLT-NAACL 2004*, Boston Massachusetts, USA, 2004, pp. 257–264.

[26] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 563–582, April 2001.

[27] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[28] 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.

[29] H. Hirsch, "FaNT-Filtering and Noise Adding Tool," 2005, http://dnt.kr.hsnr.de/download.html.