Improving Speech Synthesis for Noisy Environments

Gopala Krishna Anumanchipalli, Prasanna Kumar Muthukumar
Udhyakumar Nallasamy, Alok Parlikar, Alan W Black, Brian Langner

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, USA
{gopalakr,pmuthuku,unallasa,aup,awb,blangner}@cs.cmu.edu

Abstract

Speech Synthesizers have traditionally been built on carefully read speech that is recorded in studio environment. Such voices are suboptimal for use in noisy conditions, which is inevitable in a majority of deployed speech systems. In this work, we attempt to modify the output of the speech synthesizers to make it more appropriate for noisy environments. Comparison of spectral and prosodic features of speech in noise and results of some conversion techniques are presented.

Index Terms: speech synthesis, speech in noise, companding,

1. Introduction

As speech synthesis becomes more common we begin to care more about its performance in context and not just in test situations. In the real world there are many different forms of background noise and as humans adapt to the noise situation they are in, speech synthesis output should also do such adaptation. Such adaptation should not just be changing the gain. Humans modify their speech on a number of different dimensions in noise, including prosodic modification and spectral modification.

This paper looks at a number of different methods that allow automatic modification of speech in order to improve understandability in noisy situations. First we investigate the ANSI standard for Speech Intelligibility Index (SII) [1] and use bosting in specific frequency bands to try to improve intelligibility. We also investigate the differences between natural plain speech and natural speech in noise and explore automatic modification techniques that better model how humans modify their speech. We show the effects of both spectral and prosodic modification.

2. Speech in Noise Database

We use the CMU_SIN database [2], a database of speech in noise designed for speech synthesis. This database uses a short recording (under one minute) of human conversational babble from a crowded cafeteria to provide a noisy environment for recording, though the volume was adjusted to be clearly noticeable to the listener without being uncomfortable. The babble was played to the voice talent through headphones, along with their own speech. This simulated the acoustic environment that would actually be experienced in a noisy cafeteria, while keeping the noise out of the speech recordings. The noise was played only during delivery of the prompts, which limited the overall exposure of the voice talent to the noise, and helped to "reset" the perceived noise level in between utterances.

Since people generally adapt their speech to the conditions they are in, we cannot simply play noise to the voice talent for every prompt if we want to get a consistent elicitation of speech in noise. For this purpose, noise and non-noise conditions are randomly switched while recording. The result of this method is that during recording the voice talent was unaware of the noise condition for a particular prompt until delivering it, and seemed to consistently and appropriately produce natural speech in noise.

The transcript is a subset of the CMU ARCTIC [3] prompts for building voices; specifically, the first 500 utterances (the "A" set). Recording was done in a quiet room with a laptop, using a head-mounted close-talking microphone. Each of the 500 prompts were recorded twice, once in noise conditions and once not in noise. This was done using two separate sessions: in the first session, approximately half of the prompts were recorded in noise and half not in noise through the method described above; in the second session, the noise condition was reversed so that prompts previously recorded in noise were recorded without noise, and vice-versa. Two male speakers, one with an American English accent and one with British English accent, were recorded.

3. Measures of intelligibility in noise

The American National Standards Institute defines a measure for the intelligibility of speech in the presence of adverse conditions called the Speech Intelligibility Index (SII) [1]. This measure is believed to be correlated to the actual intelligibility of speech in such conditions. The standard states that the SII may be interpreted as a measure of the total number of speech cues available to the listener. An SII of 1.0 indicates that all speech cues reach the listener while an SII of 0.0 means the total loss of speech cues.

The index involves the computation of the 'audibility' in each band of a set of frequency bands. Each of these bands is assigned a fixed value of Band Importance. The SII is the sum of the audibility functions of these bands weighted by the band importance assigned to each band (shown in Eqn. 1),

$$S = \sum_{i=1}^{n} I_i * A_i \tag{1}$$

where I_i is the band importance function, A_i is the band audibility function and n is the number of bands used in the standard. The standard defines the band audibility function as a number between 0.0 and 1.0 that specifies the effective proportion of speech dynamic range within the band that contributes to intelligibility.

The band importance function depends on the application that this measure is used for. The accuracy of the SII is dependent on picking the right importance function. Though the SII standard defines importance functions for different tasks, for

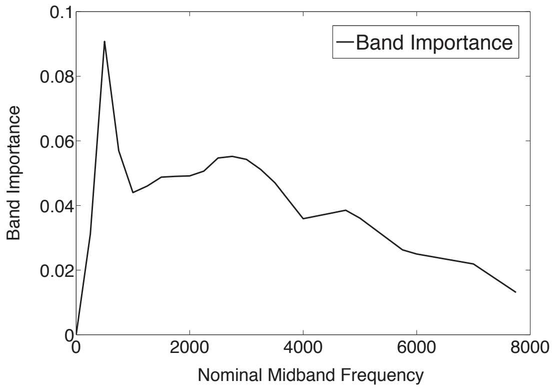

Figure 1: Band importance function for Intelligibility

the problem discussed here, we selected the function designed for intelligibility tests on short passages of easy reading material [1]. Figure 1 shows the relative importance of each frequency range for with respect to intelligibility in the presence of noise for a task involving short passages of easy reading material.

3.1. Correlation of SII with perception

In order to determine the extent to which listeners can distinguish between the two conditions, a perceptual *ABX* test is conducted where listeners were to identify which of the two presented stimuli was spoken in the presence of noise. 10 example stimuli were taken from both speech in noise and without noise databases. All of 5 subjects who took the test identified the correct instance with 100% accuracy.

The results were similar even for the synthesized examples from these databases, showing that there are perceptually identifiable changes that occur in the speech when spoken in noise. We also conducted a preference test for these stimuli. The results are summarized in Table 1 below.

#Test case	Mean preference
SIN Orig Vs. SWN orig	62.95 %
SIN Synth Vs. SWN Synth	53.33 %

Table 1: Preference scores for speech with added noise

These results suggest that subjects preferred speech recorded in noise for natural speech but not significantly so for synthetic speech. To test how these perceptual results correlate with the SII standard, we computed the SII for these examples. Table 2 shows the mean SII measures for all the above testing stimuli.

#Stimulus type	SII mean	SII Std. Dev.
SWN orig	0.229	0.027
SIN orig	0.235	0.032
SWN synth	0.284	0.039
SIN synth	0.281	0.043

Table 2: Intelligibility measures for different stimuli

For natural speech, the SII values are higher for speech in noise, in line with perceptual preference. However for synthesized examples, the difference in SII values is insignificant, and in fact is slightly greater for speech without noise. This is not correlated with the perceptual evaluations and suggests that there is a need for metrics more suitable for measuring intelligibility of synthetic speech.

4. SII-based improvements for Speech in Noise

In this section, we report experiments conducted to increase the intelligibility of normal speech in added noise. We describe two techniques to improve synthesis quality for noisy environments. The first is a post processing technique that boosts the amplitude in the important frequency bands as described by the SII standard. The second technique operates at the training stage, where each training frame is weighted with its SII value during clustering. These techniques are detailed in the sections below.

In the following sections we describe techniques that use the SII to improve intelligibility. Despite our earlier results showing low correlation between SII and perceptual preference, we believe that these techniques can give better results with measures more suited to synthesized speech.

4.1. Band significance based boosting

Since the SII is known to be highly correlated to the actual intelligibility of speech in adverse conditions, naively optimizing for this measure seemed reasonable. This involved reshaping the spectral characteristics of the speech such that the bands that were assigned higher values by the band importance function are boosted relative to bands assigned lower values. The band importance function is a relatively smooth function and so designing a filter whose magnitude response matched the function was not difficult. The filter was chosen such that the boosting that was done did not cause a significant amount of distortion. This is due to the fact that the SII is not a perfect measure of speech intelligibility and therefore, excessive optimization for the SII would not yield good results in perceptual experiments.

A subjective test was then conducted where listeners were asked to choose between the boosted synthesized speech and unmodified synthesized speech. In nearly all cases, the unmodified synthesis was judged to be more intelligible than the boosted synthesis (with added noise). While it may seem counterpointed that the boosting seems to affect intelligibility, the cause of this degradation is simple to explain. The band specific boosting applied to the speech signals worked by injecting additional power to bands that were deemed more important by the band importance function. As a result of this, the boosted speech signal overall had higher amplitude than the unmodified signal. To be able to do a valid comparison of the boosted synthesis and the unmodified synthesis, the relative amount of noise added to the synthesis needed to be the same and therefore both synthesized signals were normalized on amplitude. Since the dynamic range of the amplitude of the boosted synthesis signal was greater, normalizing this signal meant that regions of lower amplitude were suppressed greater than they were in the unmodified signal. As a result of this, the normalized boosted synthesis tended to sound quieter. Similar results were obtained when power normalization was done instead of amplitude normalization. Listeners also mentioned that they preferred the unmodified synthesis because it sounded less ‘processed’.

4.2. SII weighted clustering

This section describes our experiment on computing Clustergen [4] codebooks using a SII-based weighting scheme. Our approach is similar to [5] where SII measure is used in a unit selection setup. The intelligibility cost is optimized in addition to unit cost and join cost in choosing a particular unit for concatenation. We extend the concept of SII cost function to parametric synthesis. In Clustergen, contextual decision trees are used to cluster speech frames in each labeled state into various representative clusters. A single Gaussian is estimated for each cluster. The first and second order sufficient statistics are calculated as follows.

$$\theta_c(\boldsymbol{o}) = \sum_{i=1}^{N_c} x_i, \qquad \theta_c(\boldsymbol{o^2}) = \sum_{i=1}^{N_c} x_i^2 \tag{2}$$

where N_c is the number of speech frames in each cluster and x_i is the feature vector. The cluster-specific mean and diagonal covariance are then computed as

$$\mu_c = \frac{\theta_c(\boldsymbol{o})}{N_c}, \qquad \Sigma_c = \frac{\theta_c(\boldsymbol{o^2})}{N_c} - \mu_c^2 \tag{3}$$

If γ_i is the SII value for frame i in cluster c. From the definition of SII, $0 \leq \gamma_i \leq 1$. The count and SII-weighted sufficient statistics are then calculated as

$$\gamma_c = \sum_{i=1}^{N_c} \gamma_i, \quad \hat{\theta}_c(\boldsymbol{o}) = \sum_{i=1}^{N_c} \gamma_i x_i, \quad \hat{\theta}_c(\boldsymbol{o^2}) = \sum_{i=1}^{N_c} \gamma_i x_i^2 \tag{4}$$

The cluster-means and diagonal covariances are

$$\hat{\mu_c} = \frac{\hat{\theta}_c(\boldsymbol{o})}{\gamma_c}, \qquad \hat{\Sigma_c} = \frac{\hat{\theta}_c(\boldsymbol{o^2})}{\gamma_c} - \hat{\mu_c}^2 \tag{5}$$

For each training speech utterance, we generate a white noise waveform with equal power. The silences within the speech waveform are excised using a simple threshold function to improve the reliability of frame-wise SII score [6]. The SII values are then calculated for every 25ms frame with a 5ms shift [7]. A Hamming window is used to smoothen the edges before SII computation. The frame and shift sizes are chosen to coincide with MCEP feature extraction for easy manipulation of SII scores. CLUSTERGEN parametric synthesizer is built following the usual steps, except that the cluster-specific means and covariances are calculated using the weight adjusted formulae shown above. There are no modifications to the synthesis procedure. We used Roger Arctic set [8] with 1132 utterances for this experiment. Every 10th utterance is pooled into a held-out test set (113 utterances) and the rest (1019 utterances) are used as training set.

For objective testing, we calculated SII values for each synthesized waveform and computed the mean SII for the entire test set. The results are shown in Table 3.
As shown, SII-weighted clustering performs better than the baseline method with respect to the SII metric. Thorough subjective tests are required to further confirm this result.

It is important to note that although the SII measure is used to predict the intelligibility of speech in noise, it is not very robust when calculated within segments of duration of a phone or state. There have been techniques proposed in the literature to improve the reliability of SII within short segments [9], which is now an addendum to the original ANSI-S3.5 SII standard. We plan to experiment with these improved intelligibility measures in future.

codebook type	SII-Mean	SII-Stdev
baseline	0.4590	0.0948
sii-weighted	0.4711	0.0941

Table 3: SII values for synthesized speech

5. Spectral analysis of speech in noise

In order to study any systematic changes in the spectral characteristics of speech in the presence of noise, we analyzed the speech delivered in both conditions. We used the database described in a section 2. Appendix A shows the spectrograms of the same sentence spoken under within and without the presence of noise.

As can be observed, segments of speech change differently in the presence of noise. There is no consistent change in characteristics like the sub-band energies of different speech regions. To study the spectra in more detail, the average spectral behavior of each phoneme was analyzed. The Fourier transforms of speech segments are used as the spectral features. For each 5 ms of speech, 512 point FFTs are extracted. The phonetic labels are used to accumulate the statistics of all the frames that belong to a particular phoneme. To isolate the behavior of the phoneme from reticulation effects, only the ‘middle’ states of the phonemes are used. This is assuming that the middle state roughly corresponds to the steady state of the phoneme. The mean spectra of phonemes are compared in the two conditions.

Appendix B shows the mean FFT spectra of a example voiced and unvoiced phonemes over all of the utterances. We observed that for vowels there is a pronounced increase in the formant amplitude in the mid-frequency regions (3000-5000 Hz, i.e, mainly the second and third formants frequencies). For other phonetic categories, the difference is almost negligible with some exceptions. The interesting find is that the relative changes in the spectra of all phonemes are similar within two English speakers. Similar analysis done on a Mandarin speech corpus gave results that were not comparable to English, suggesting a speaker-independant, language specific behaviour of speakers in noise. The research on Mandarin is preliminary at this stage and not reported further in this paper.

Based on the phoneme specific differences observed above, we designed FIR filters to modify the spectral properties of speech without noise to match that of speech in noise. However, the perceptual evaluations with added noise did not show improvemts, partly due to the normalization issues previously described in Section 4.

6. Prosodic analysis of speech in noise

Complementing the spectral analysis and conversion techniques presented above, this section addresses the changes in prosodic aspects observed within speech in noise.

6.1. Duration

Duration is a key prosodic aspect that speakers change when speaking in noise. The goal of this study is to modify durations of a normal synthetic voice to suit the noisy environment conditions better.

The same speech database as in Section 2 is used for this analysis. The speaking rate of the speaker is measured under each condition. We use the definition of speaking rate to be the number of syllables uttered per second. The mean and the standard deviations are reported in Table 4. The mean speaking rate are not significantly different for speech spoken without noise, but the standard deviation is much lower. This could be probably because of the fact that speakers use speaking rate as a communicative device and have greater degree of freedom in the noise-free recording condition. While speaking in noise, to assist the listener, speakers may be compensating for the noise by speaking consistently with lesser variation.

#Background	Mean speaking rate	Std. dev.
without noise	4.89	1.046
in noise	4.857	0.669

Table 4: Speaking rates of speech in clean/noisy backgrounds

6.1.1. Statistical modeling of duration

On an average, there were a larger number of phonemes whose duration is longer in speech recorded in the presence of noise than in speech recorded in the absence of noise. However, the relative proportion of phonemes with longer durations is not substantial enough to make any conclusions. Only 66% of the phonemes had longer durations while the rest had a shorter duration. Patel et al. [10] approache the problem of duration by heuristically increasing the duration of content words with an average stretch of 98ms. Extending this, we investigated the role of Part of Speech in determining the change in duration of the word. This analysis of across a large number of utterances did not suggest any consistent pattern. Some phonemes were on average longer while others were on average shorter. There was, however, a large variance in the durations of all phonemes.

To learn these changes automatically from data, we built decision trees to determine the stretch/shrink factor based on context features. The duration model consists of means and standard deviations of the durations of each HMM state and a decision tree is used to obtain the corresponding $z-$scores.

Using the above mentioned duration modeling technique, the following were attempted –

1. Use a duration model with z-score CART from speech-in-noise, and means from speech-without noise.
2. Use a duration model with z-score CART from speech-without-noise and means from speech-in-noise.
3. Use the duration model with both z-score and means from speech-in-noise, with every other model in the speech-without noise condition.

In all three cases, listening tests revealed that there was no significant improvement in intelligibility (with added noise) despite a noticeable change in phoneme durations.

6.2. Fundamental frequency (F0)

Similar analysis was conducted on $F0$s of the speaker under both conditions. Table 5 shows the means and standard deviations of F0 under the two conditions. A significant increase in the pitch is observed for speech in noise.

#Background	Mean F0	Std. dev.
without noise	126.953	34.450
in noise	136.254	33.091

Table 5: F0 of speech in clean/noisy backgrounds

A z-score normalization procedure, similar to the duration modeling experiment can be done to account for the F0 raise.

It remains to be seen how effective the prosodic modification techniques perform when done in tandem with spectral modifications.

7. Post-processing in the time domain

As noted earlier, Sections 4 and 5 had the recurring problem of normalizations when adding noise. Changing the spectral characteristics of speech, in all cases, resulted in an increase in the dynamic range of the amplitude of the signal. Perceptual experiments had yielded the result that, after amplitude normalization, the speech signal with the larger dynamic range sounded quieter and therefore harder to understand in the presence of noise. It therefore seems that the dynamic range of the amplitude of the speech signal plays an important role in how it is perceived. Lowering the dynamic range makes the signal sound louder and therefore easier to understand in adverse conditions. However, this decrease in dynamic range also introduces distortions that make the speech signal sound noisier. Therefore, there is a tradeoff between loudness and distortion when changing the dynamic range.

A rather naïve approach to improve the intelligibility of speech in noise, the μ-law algorithm performed the most effectively. This method is known as the *compand* operation (compression + expansion) on the speech signal. This has the effect of making the quieter sounds in speech sound louder while affecting louder sounds to a lesser extent. The μ-law algorithm is a standard technique used to lower the dynamic range of an audio signal for various purposes. However, this algorithm is primarily used for audio coding and using parameters of similar magnitude would cause a substantial amount of distortion. The value of the μ parameter in the μ-law coding algorithm was chosen experimentally rather than trying to optimize for any particular intelligibility measure.

8. Conclusion

This paper has identified a number of differences between natural speech and natural speech in noise. It has investigated both objective measures and tested with subjective measures. Although natural speech in noise is does reflect improvements in the objective SII measure and in listening tests, speech modification techniques do not give such a clear benefit. The only technique we have found so far that gives a significant improvement is the somewhat simple companding techniques, even though that techniques ignores all of the identified spectral, and prosodic differences between the two types of speech.

This work implies there are still more subtle aspects of speech in noise that we are not yet modeling properly for improving the intelligibility of synthetic speech in noise. An integrated approach to joint spectral and prosodic modifications for this problem is still due.

9. Acknowledgements

The first author is supported by the Fundação de Ciência e Tecnologia through the CMU/Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.

10. References

[1] *Methods for Calculation of the Speech Intelligibility Index ANSI S3.5-1997*, American National Standards Institute, 1430 Broadway, New York, NY 10018, USA, 1997.

[2] B. Langner and A. Black, "Creating a database of speech in noise for unit selection synthesis," in *5th ISCA Workshop on Speech Synthesis*, Pittsburgh, PA, 2004.

[3] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu_arctic/, 2003.

[4] A. Black, "Clustergen: A statistical parametric synthesizer using trajectory modeling," in *Interspeech 2006*, Pittsburgh, PA, 2006.

[5] M. Cernak, "Unit selection speech synthesis in noise," in *ICASSP 2006*, Toulouse, France, 2006.

[6] K. D. Donohue, "Audio systems array processing toolbox." [Online]. Available: http://www.engr.uky.edu/ donohue/audio/Arrays/MAToolbox.htm

[7] ASA, "Speech intellegibility tools." [Online]. Available: http://www.sii.to/siimatlab.zip

[8] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Blizzard Challenge 2008*, Brisbane, Australia, 2008.

[9] K. S. Rhebergena and N. J. Versfeld, "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[10] R. Patel, M. Everett, and E. Sadikov, "Loudmouth:: modifying text-to-speech synthesis in noise," in *Assets '06: Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. New York, NY, USA: ACM, 2006, pp. 227–228.