IMPROVING THE UNDERSTANDABILITY OF SPEECH SYNTHESIS BY MODELING SPEECH IN NOISE

Brian Langner, Alan W Black

Language Technologies Institute
Carnegie Mellon University
{blangner,awb}@cs.cmu.edu

ABSTRACT

Although the quality of synthetic speech has increased dramatically in the past several years, many people still have difficulty understanding speech produced by even the highest quality synthesizers. We describe an approach to improve understandability of synthetic speech using *speech in noise*. Natural speech in noise is a change in the style of speech that is used by people to improve the understandability to the listener when speaking in poor channel conditions. We show that altering the presentation of synthetic speech in similar ways also improves understandability. Further, we discuss methods of obtaining speech in noise for use in speech synthesis, as well as the results of an evaluation of several synthetic voices that "speak in noise".

1. BACKGROUND

Despite vast improvements in the quality of synthetic speech, many people still find it difficult to understand, even when the best synthesizers are used. The CMU Let's Go! project [1] is developing techniques to improve spoken dialog systems for non-native speakers and the elderly; specifically, improving the quality of spoken output to make it more understandable by those groups, as well as the general population. There are a number of factors that have an impact on understandability, including lexical choice, prosody, and spectral qualities of the speech itself. In an earlier experiment which used recorded natural speech [2], it was found that understandability improved when the speech was delivered as if the listener had said, "I can't hear you, can you say that again." This change in speaking style can be elicited from people by having them speak in a noisy room.

In order to reliably elicit such a delivery style – speech spoken in poor channel conditions – as well as obtain clean recordings for use in speech analysis and synthesis, we used the method that produced the CMU SIN database for speech synthesis [3] to record a small (30 sentence) database of speech in noise. The CMU SIN database is publicly available at `http://festvox.org/cmu_sin/`.

The smaller database used in this work was recorded in a quiet room with a laptop and head-mounted close-talking microphone. During recording, the voice talent wore headphones, which played both noise and the voice talent's own speech to simulate a noisy room while still providing audio feedback of the speakers voice. In this way, we were able to obtain clean (noiseless) recordings suitable for use in concatenative speech synthesis, while simultaneously providing an environment to the voice talent that seemed quite noisy. The noise used during the recording process was a short recording of a crowded cafeteria during peak lunch hours, a type of noise selected because of its naturalness, people's familiarity with it, and the ease with which it can be obtained.

Unfortunately, people generally will adapt their speech based on the conditions they are in, so we cannot simply play noise to the voice talent for every prompt if we want to get a consistent elicitation of speech in noise. Thus, for each utterance, we randomly played noise (or not) while the voice talent spoke, with the intention of not allowing the speaker to become too accustomed to the noise. We added the condition that no more than three consecutive prompts would be recorded in the same noise / non-noise condition, to ensure that even in the short term, the voice talent would not be able to adjust to the noise too much.

However, since we cannot record all the prompts in noise at once, the result is that two recording sessions are required to build a database of speech in noise, with the in-noise and not-in-noise conditions reversed for each session, giving us an identical database of plain speech in addition to the database of speech in noise. The content of this database consists of sentences describing times when public buses in Pittsburgh leave specific stops. The "bus information" domain is large, but finite, and the subset we are using here is relatively constrained; some statistics on this domain are shown in Table 1.

It should be noted that speech in noise is not *only* louder than plain speech; it has different spectral qualities, different prosody, and different durations. Such speech has sometimes been referred to as Lombard speech [4], but we do not

Sentences	Words	Phones	Buses	Stops
30	418	1464	7	12

Table 1. The number of various units and concepts used in the bus information domain database.

feel that term is appropriate for this work, as the level of background noise being used here is fairly small. Also, we are not working with more extreme examples of speech in noise, such as shouting.

2. EVALUATING NATURAL SPEECH IN NOISE

We first wanted to confirm that the results of [2] were applicable to the style of speech we obtained in our speech in noise database. Thus, using the natural recordings of plain speech and speech in noise in the bus domain, we designed a listening test for understandability. Since natural speech is generally easy for most people to understand, we chose to add noise to the recordings to increase the difficulty of the task. The noise source was again human conversational babble from a crowded cafeteria. Three noise conditions were used: no noise (original recordings), added noise giving a -3.2 dB signal-to-noise ratio, and added noise giving a -4.9 dB signal-to-noise ratio. The signal-to-noise ratio was calcuated using the ratio of the average power of the sentences and the power of the added noise; the formula is shown here:

$$SNR = 10 \times \log_{10} \frac{P_{signal}}{P_{noise}}$$

Because the power of the noise we are adding is greater than the power of the speech (the ratios are 48% and 32%, respectively), the resulting signal-to-noise ratios are negative. These noise levels were chosen based on the results of an empirical study. People were asked to listen to recordings with a wide range of signal-to-noise ratios (from -12.1 dB to +8.7 dB).

The content of the recordings was identical in all cases: a single natural, plain speech sentence of bus information. -3.2 dB is the signal-to-noise ratio at which most people were able to get some words in the sentence while still making some errors, whereas a ratio of -4.9 dB is the level at which few people could understand more than a couple of words. If speech in noise were more understandable under poor channel conditions, we would expect the speech in noise recordings to have fewer errors than the plain speech recordings.

Although power alone is a important factor in understanding speech, we wish to test the effect of other dimensions, such as pitch, duration, and spectral shape. Because speech in noise is, on average, louder than plain speech, in order to ensure that power differences alone do not account for any observed improvements, we normalized the power of all the recordings to the average power of the plain speech recordings.

To evaluate the relative understandability of speech in noise, we had ten people listen to four examples of each speaking style / noise level combination, for a total of 24 sentences. These sentences were arranged randomly, with the stipulation that the same condition could not be heard twice in a row. Listeners were asked to listen to the sentences as few times as possible (generally, this was three or fewer), and type in all of the words in the sentence that they could understand. These were then scored using word error rate (WER). The results are shown in Table 2.

It is clear from these results that speech in noise is easier to understand than plain speech when the noise level is high. This result is independent of the typical power differences between plain speech and speech in noise, as well, because of the power normalization we have done. This suggests that the spectral, prosodic, and durational differences have a positive influence on understandability in noisy conditions. There does not seem to be a significant effect on understandability when conditions are not noisy, or even moderately noisy, however. This is likely because under "easy" conditions, people are generally able to understand natural speech.

3. MODIFYING OTHER VOICES

While concatenative unit selection synthesis is capable of producing high-quality synthetic speech, there are limitations to this technique; it depends on the existence of suitable examples within the database to select from. Thus, when we require different speaking styles, we must record these new styles in separate databases [5]. Since we would like to be able to use improvements in understandability with existing synthetic voices, not just newly created ones, we require models of speech in noise that we can apply to produce voices that speak in noise without necessitating extra recording.

A number of factors distinguish plain speech from speech in noise. Conventional speech synthesis prosody modifica-

Style	S/N	Avg. WER
Plain	no noise	0.19%
In Noise	no noise	0.98%
Plain	-3.2 dB	5.83%
In Noise	-3.2 dB	6.82%
Plain	-4.9 dB	25.98%
In Noise	-4.9 dB	12.15%

Table 2. Word-error-rate scores for plain speech and speech in noise at various signal-to-noise (S/N) ratios.

tions such as pitch, power, and duration are important factors, but the differences between plain speech and speech in noise go beyond just those. We also wished to investigate how spectral differences affect understandability of speech. Using techniques that were designed for voice conversion between a source and target speaker [6], we applied such techniques to style conversion to learn a mapping between plain speech and speech that was generated in noise. This work uses a Gaussian Mixture Model (GMM) transformation method [7], as distributed with the FestVox tools [8].

Two different modification models were built. The first was trained between the in-noise / not-in-noise data collected in this experiment. The resulting model was then applied to an existing domain targeting unit selection voice, recorded by the same speaker, that was built for the Let's Go! project. The second voice we tested was a standard diphone based voice – the "kal_diphone" voice from the Festival Speech Synthesis System [9]. In this case, we trained a model that converted kal_diphone to our speech in noise databases; this conversion involves two different speakers.

4. EVALUATING MODIFIED VOICES

To evaluate the effectiveness of our modified synthetic speech, we used a similar process as with the natural speech in noise. Again, to account for power differences, all of the samples were power normalized to the level of natural plain speech. To increase the difficulty of the task, we added noise to the sentences as before, producing noise conditions with signal-to-noise ratios of -3.2 dB and -4.9 dB, as well as no noise. Two different synthetic voices were used: a diphone voice and a unit selection voice built for this domain. Furthermore, both voices were also modified using the style conversion process described above, for a total of four different voice conditions.

The same ten listeners from above were asked to listen to three examples of each voice / noise level condition, for a total of 36 sentences. Again, they were directed to listen to each sentence as few times as possible, and type in all of the words they were able to understand. The word error rate results are shown in Table 3.

There are several things to note from these results. First, the modified diphone voice shows a dramatic improvement in understandability under moderately noisy conditions, with a 25% absolute reduction in word error rate. Even under higher noise conditions, the modified voice is more understandable, though the difference is not nearly as great. However, given the high error rates at that noise level, it is likely that the content was simply drowned out by the noise. Since the domain is predictable for people with knowledge of the bus system in Pittsburgh, reasonable guesses will often be correct. With no noise, the modified voice has an increased error rate, though this could be influenced by a number of different factors, such as the presence of tokens which are easily confusable (for example, the bus number "71D" has several valid, similar-sounding alternatives, such as "71B" or "71C"). Further, the style conversion process does introduce some degredation of the signal, which is noticable in good channel conditions; such degredation could exacerbate problems with confusable tokens, explaining the increased error rate.

Voice	Style	S/N	Avg. WER
Diphone	Plain	no noise	0.70%
Diphone	In Noise	no noise	2.82%
Diphone	Plain	-3.2 dB	28.38%
Diphone	In Noise	-3.2 dB	3.11%
Diphone	Plain	-4.9 dB	33.07%
Diphone	In Noise	-4.9 dB	31.43%
Unit Sel.	Plain	no noise	1.19%
Unit Sel.	In Noise	no noise	1.36%
Unit Sel.	Plain	-3.2 dB	2.20%
Unit Sel.	In Noise	-3.2 dB	9.33%
Unit Sel.	Plain	-4.9 dB	8.73%
Unit Sel.	In Noise	-4.9 dB	10.92%

Table 3. Word-error-rate scores for different synthetic voices at various signal-to-noise (S/N) ratios.

Second, the modified unit selection voice does not show any improvement over the unmodified version, and in fact shows a significant decrease in understandability with moderate noise. One possible reason for this is the distortion introduced by the signal processing in the conversion. The converted speech is reconstructed from cepstral vectors using a vocoder which reduced the overall quality of the signal. Any advantage that may be given by the speech in noise modification is apparently lost by the signal processing. The positive diphone result may explained by the fact that the diphone quality, from residual excited LPC, is not all that much different from the vocoder quality output of the style converted voice.

5. CONCLUSIONS AND FUTURE WORK

We have confirmed that natural speech in noise can improve understandability of speech delivered under poor channel conditions. We have determined that the increase in understandability is not solely due to the power differences between speech in noise and plain speech, but is also affected by the spectral, prosodic, and durational differences between the speech styles.

By applying voice conversion techniques, we have demonstrated that it is possible to modify existing synthetic voices to speak in noise if suitable databases to train a mapping

between plain speech and speech in noise are available. Using this style conversion, we have shown that a diphone voice can have its understandability significantly improved for noisy channel conditions.

Our evaluation of speech in noise used sentences in a constrained, and thus predictable, domain. While the use of this domain does provide a real-world task in which to test our voices, its predictability means that people are able to guess the correct word or words in a sentence when they did not understand it well. One possible solution to this problem would be to select content from the domain that participants in the evaluation are unfamiliar with, such as stops and routes from far outlying areas rather than neighborhoods located near the universities, which should make it more difficult to "guess" correctly when a difficult word is encountered. Another solution would be to use semantically unpredictable sentences [10] instead of the domain sentences used here. These sentences would follow a specific syntactic pattern, such as "Determiner Adjective Noun Verb Determiner Adjective Noun.", but be filled with words whose juxtaposition is unlikely. Using sematically unpredictable sentences would also allow us to perform a domain-independent evaluation of understandability, so that people who are more familiar with the domain will not have an advantage when guessing about hard-to-understand words.

In addition, though the improvement shown by a modified diphone voice is encouraging, a speech in noise style conversion must also work for unit selection voices to be useful. There is room for improvement in the plain speech to speech in noise mapping for the unit selection voice, which would result in a higher quality unit selection voice that speaks in noise.

6. ACKNOWLEDGEMENTS

This work is supported by the US National Science Foundation under grant number 0208835, "LET'S GO: improved speech interfaces for the general public". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] A. Raux, B. Langner, A. Black, and M. Eskenazi, "LET'S GO: Improving spoken dialog systems for the elderly and non-native," in *Eurospeech*, Geneva, Switzerland., 2003.

[2] M. Eskenazi and A. Black, "A study on speech over the telephone and aging," in *Eurospeech01*, Aalborg, Denmark, 2001.

[3] 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.

[4] H. L. Lane and B. Tranel, "Le signe de l'elévation de la voix," *Annales Maladiers Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[5] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to <AHEM/> expressive speech synthesis authors," in *5th ISCA Workshop on Speech Synthesis*, Pittsburgh, PA, 2004.

[6] T. Toda, *High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion*, Ph.D. thesis, Nara Institute for Science and Technology, 2003.

[7] Y. Stylianou, O. Capp'e, and E. Moulines, "Statistical methods for voice quality transformation," in *Proc. EUROSPEECH95*, Madrid, Spain, 1995, pp. 447–450.

[8] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," http://festvox.org/bsv/, 2000.

[9] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," http://festvox.org/festival, 1998.

[10] C. Benoit, M. Grice, and V. Hazan, "The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences.," *Speech Communication*, vol. 18, pp. 381–392, 1996.