CREATING A DATABASE OF SPEECH IN NOISE FOR UNIT SELECTION SYNTHESIS

Brian Langner, Alan W Black

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

ABSTRACT

This paper describes CMU_SIN, a new database of speech in noise that can be used for unit selection speech synthesis. We describe a process that can be used to elicit speech in noise and how to use that as part of building a synthetic voice that speaks in noise. Details of the database we constructed, as well as some preliminary analysis and future goals of this work, are also included.

1. INTRODUCTION

As we move to higher quality synthetic speech through unit selection techniques, we also lose some control of the speech signal. Unit selection concatenative synthesis can produce high quality speech, but 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. With an aim to better model different styles rather than require full recordings, we have designed and built a database that addresses speech output in noise.

Within the CMU Let's Go project [1], we are developing techniques to improve spoken dialog systems for non-native speakers and the elderly. Specifically, we wish to improve the quality of spoken output to make it easier to understand. There are a number of factors that affect understandability including lexical choice, prosody, and spectral qualities of the speech itself. In an earlier experiment [2] where we used recorded natural speech, we noted that understandability improves when the speech was delivered as if the listener had said, "I can't hear you, can you say that again."

In order to investigate such a delivery style – speech spoken in poor channel conditions – we have designed and recorded a database that captures the style so that we might better model it and be able to apply it to other voices. It should be noted that speech in noise is not simply louder; it has different durations, a different tune, and a different spectral quality. Such speech has sometimes been referred to as **Lombard speech**, but we refrain from using that term as the level of background noise we are using is fairly small. We are not, at this stage, looking for more extreme examples of speech in noise such as shouting.

2. BUILDING VOICES IN NOISE

Building voices that capture qualities of speech in noise requires some modification to a typical synthetic voice building process, such as the Festvox voice building tools [3]. Obviously, while the voice talent is recording the prompts for the voice, it is necessary to have an audible noise source altering the talent's delivery. We used a short (less than a minute long) recording of human conversational babble from a crowded cafeteria. This provided an easily obtainable, "natural" noise condition to speak in that most people should be familiar and comfortable with.

We used this babble to provide a noisy environment for recording, first adjusting its volume so it would be clearly noticeable to the listener without being uncomfortable. The babble was played to the voice talent through headphones, along with the talent's utterances, simulating the acoustic environment that would actually be experienced in a noisy cafeteria, while keeping the noise out of the recording of their speech. 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. Adding this to the voice building process required us to modify the recording script to play from a sound file at the same time as the voice talent delivered a prompt.

However, since people generally will 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 reason, we decided to randomly switch between noise and non-noise conditions while recording. Our modifications to the recording script thus also included a mechanism to randomly choose to play noise as the prompt was recorded, with the stipulation that no more than three consecutive prompts would be recorded in the same condition. The latter condition is designed to ensure that even in the short term, the voice talent would not be able to adjust to the noise too much. 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.

We used a subset of the CMU ARCTIC [4] prompts for building these voices; specifically, the first 500 utterances (the "A" set). Some statistics about the number of units in

Prompt Set	# Prompts	# Words	# Phones
CMU_SIN	500	4414	17322
ARCTIC "A"	593	5284	20677
Full ARCTIC	1132	10045	39153

Table 1. The number of various units in the prompt set used for this database, compared to CMU ARCTIC prompt sets.

this prompt set, as well as comparisons to relevant ARCTIC sets, are shown in Table 1. These prompts were selected because they provided a reasonably large, phonetically balanced data set to record from, while being a sufficiently small set that they could be recorded relatively quickly without overtaxing the voice talent. Furthermore, the voice talent had previously recorded one of the distributed ARCTIC voices, and so was already somewhat familiar with the prompts.

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. This allowed us to produce two comparable voices: a voice that "speaks in noise" and an otherwise equivalent baseline. The data was automatically labeled with speaker specific acoustic models, no hand-correction has been made, and full unit selection synthesizers were built.

3. DISCUSSION AND FUTURE WORK

It is clear that there are properties of speech in noise that distinguish it from "normal" speech. For a person listening to the speech from the two voices we created, it is trivial to recognize which voice was recorded in noise and which was not. In addition to the audible qualities that are human-detectable, duration models trained on these databases select different features depending on whether the recordings were done in noise or not. This suggests that the properties of speech in noise that differ from normal speech can be detected by machine as well.

The difference in the length of recorded speech between the voices, though it is not very large (only 3% over 21 minutes), seems to suggest that the speech in noise is, in general, slower than normal speech. Examining the mean phoneme durations obtained from the models we trained, the durations for the voice in noise are, on average, about 3.3% greater than those for the baseline voice, confirming that speech in noise tends to be slower than normal speech.

We plan to perform some tests to evaluate the effectiveness of the voice in noise. One possibility, following [5], would be to synthesize several semantically unpredictable sentences with both the voice in noise and the voice not in noise. 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. We would then add varying levels of conversational babble to the resulting waveform files, and have subjects write down the sentence they heard; the word error rate should provide some insight into the intelligibility of these voices. It may also be interesting to use different kinds of noise as well, to determine if the speech is adapted to a specific kind of noise, or whether it is understandable in many kinds of noise.

Should the voice prove to be understandable under noisy conditions, we hope to model the appropriate qualities that differentiate speech in noise from other speech, and apply a similar method as in [6] to use this model with other general-purpose voices. If successful, we would then have a method for providing voices that can speak in noise without requiring that a specific "in noise" database be recorded.

The database is available at
`http://www.festvox.org/cmu_sin`.

4. ACKNOWLEDGMENTS

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.

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

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

[5] 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.

[6] A. Raux and A. Black, "A unit selection approach to f0 modeling and its application to emphasis," in *ASRU2003*, St Thomas, Virgin Is., 2003.