USING SPEECH IN NOISE TO IMPROVE UNDERSTANDABILITY FOR ELDERLY LISTENERS

Brian Langner, Alan W Black

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

ABSTRACT

This paper describes work designed to improve understandability of spoken output, specifically for the elderly, by using a speaking style employed by people to improve their understandability when speaking in poor channel conditions. We describe an experiment that shows the understandability gains that are possible using naturally-produced examples of this style. Additionally, we describe how to model this style, and evaluate the differences in understandability for speech synthesis produced using those models.

1. INTRODUCTION

1.1. Background and Motivation

When humans are confronted with situations where their speech is difficult to understand, they will change the manner in which they produce speech in a variety of ways to improve how understandable they are. At least one experiment [1] has shown, using recorded natural speech, that people were better able to understand what was said when when the speech was delivered as if the listener had said, "I can't hear you, can you say that again." This change in delivery style can be referred to as *speech in noise* [2], or speech spoken in poor channel conditions. Speech in noise can be elicited from people by having them speak in a noisy room. In order to investigate this speaking style, we have designed and recorded a database of natural speech in noise.

It should be noted volume is not the sole difference between speech in noise and "normal", or plain, speech. Speech in noise has different spectral qualities, different durations, and different prosody than plain speech, in addition to the power differences. Such speech has been referred to as *Lombard speech* [3], but we feel that term is inappropriate for this work, because the level of background noise we are using is relatively small. Furthermore, this work does not deal with more extreme examples of speech in noise, such as shouting.

Speech in noise can have different properties depending on the type of noise the speaker is dealing with. For example, speech produced during a rock concert will be different than speech produced near a loud white noise source, and both of those will be different than speech produced in a noisy restaurant. This work uses a recording of human conversational babble from a crowded cafeteria during peak lunch times as the noise source; thus, any conclusions from this work are likely limited to similar noise sources. The noise source was selected for several reasons, including its naturalness, people's familiarity with it, its spectral qualities, and the ease with which it could be obtained. Though our findings may be applicable in other circumstances, this has not yet been shown to be true, and so this work should not be taken as authoritative for all types of speech in noise. However, the speech collection and evaluation methods we have used are relevant for most, if not all, types of speech delivery styles worth investigating, and so this work provides a possible framework for working with speech beyond the specific style detailed here.

While we are interested in the understandability effects of natural speech in noise, our interest is motivated by our ability to get similar increases in understandability for synthetic speech. Despite vast improvements in the quality of speech synthesis in recent years, many people continue to find even the highest quality synthetic speech difficult to understand. Through the CMU Let's Go! project [4], we are developing methods to improve spoken dialog systems for non-native speakers and the elderly; specifically, we are working to improve the spoken output to make it more understandable by those groups, and by extension, the general population. If we could see understandability improvements for computer-generated speech like those of natural speech in noise, applications of synthesis such as spoken dialog systems would become significantly more usable in non-research environments. Recent preliminary work [5] suggests that such improvements are possible, though not trivial, to obtain.

1.2. Speech In Noise for Speech Synthesis

It would seem at first glance that recordings of speech in noise are relatively simple to get: just have your voice talent recorded in a room with the noise source you want. While this would provide recordings of speech in noise, those recordings would be essentially useless for any kind of synthesis or evaluation task due to the background noise that would be present in the recordings along with the speech. Unlike speech recognition, where work with speech in noise requires the corresponding background noise with the speech for good results, concatenative speech synthesis as well as human perception of speech are significantly degraded if noise is present in the speech recordings. Since those are the tasks we are concerned with, we must have a way of recording the *style* of speech in noise without contaminating our recordings with noise; what we require is recordings of *clean* speech in noise. For this paper, the phrase 'speech in noise' refers to those clean examples.

Furthermore, speech databases for high-quality (concatenative) synthesis need to contain many consistent examples of the units that are combined to produce synthetic utterances. Thus, we need a relatively consistent noise source to be certain that the

recorded speech is as suitable as possible for this use. Simply recording in a noisy room, even with a way to isolate the desired speech from the noise, is not likely to be sufficient, as natural, live noise sources are rarely consistent enough over the time period required to record a database of any reasonable size. Even worse, human speakers are annoyingly adaptive, changing their speech production as they "get used" to the conditions they are in. This tends to result in prompts recorded earlier differing in style from the later prompts, leaving a database that is unsuitable for quality speech synthesis. Given these problems, it was necessary to design a recording method [6] that would account for them. The recordings used in this work were made in a quiet room with a head-mounted, close-talking microphone.

In order to isolate the desired speech from the noise source in the recordings, the voice talent should wear headphones during the recording process. The headphones deliver the noise source as well as the voice talent's own speech; effectively, this simulates the acoustics of a noisy room to the voice talent without putting the noise in the same channel as their speech. Obviously, the noise source should be pre-recorded to simplify the logistics of playing it through headphones. It should be noted that the volume of the noise source can, and should be, adjusted to the desired level; in our work, it was adjusted to a level where it was noticeable to the voice talent without being uncomfortable. This approach accounts for both isolating the speech from the noise source, as well as the consistency of the noise source, though we must still deal with the adaptability of the voice talent.

Because of that adaptability, we cannot simply play the noise source to the voice talent continuously during the recording session if we want a consistent elicitation of speech in noise. For this reason, the noise source should be played through the headphones only while a prompt is being recorded, limiting the overall exposure of the voice talent to the noise, and helping to "reset" the perceived noise level in between utterances. However, this is insufficient, as people will still adapt to the noise over the course of recording a reasonably-sized database. Therefore, the noise should be randomly played or not played while a specific prompt is being recorded, so that the voice talent is unaware of the noise condition ahead of time. Our work limited the number of consecutive prompts with the same noise/non-noise condition to three, to ensure that even in the short term, it would be difficult for the voice talent to adjust. It is unclear if this condition is strictly necessary, but our results show that we were able to elicit consistent and appropriate speech in noise from the voice talent.

This method, while producing recordings of clean speech in noise, does have its drawbacks compared to a normal process for recording a speech database. The most notable drawback is that two full passes through the database are required to obtain a single speech in noise database. The first pass records approximately half the prompts in the noisy condition, and the second pass reverses the noise/non-noise condition for the individual prompts so that each prompt is recorded with noise. This effectively doubles the required recording time. Recording in noise is also somewhat more taxing for the voice talent, so the length of a recording session is more limited than normal. However, the method does produce two parallel databases in the end – a database of speech in noise, and an otherwise identical one of plain speech – which can be useful in several different applications.

After recording a database of speech in noise, it is possible to build a voice using that data just as with any other database, with a few caveats. As noted above, speech in noise has different spectral and prosodic qualities than plain speech. This often causes methods of F_0 extraction to give poor results, which in turn lowers the quality of the resulting synthesis. Additionally, to this point our work has been done with relatively small databases (under 30 minutes of speech), which limits the attainable synthesis quality compared to what can be achieved with larger databases. However, given the extra effort involved in building a speech in noise database as compared to normal synthesis databases, and our desire to explore the usefulness of this kind of style modification, we did not feel it was worthwhile to invest the time required to record more than a small database.

Furthermore, it is possible to build a single voice that can produce plain speech or speech in noise, using marked-up text to determine the speaking style, since the recording process generates a full database of both styles. Such a voice is useful for circumstances where having multiple distinct voices is undesired or infeasible, but both speaking styles are required.

2. MODIFYING OTHER VOICES

While it is possible to produce high-quality synthetic speech in noise by building a voice out of speech in noise recordings, that method has a significant drawback: it requires recording an entirely new database for each application. Furthermore, if styles other than speech in noise are desired, each style will require its own database of recordings [7]. Clearly, this is not an ideal solution, especially for applications which already have existing synthetic voices. Since we would like to be able to make use of understandability improvements in many applications, including those which have pre-existing voices, we require models of speech in noise that can be applied to produce the style without necessitating re-recording of an entire database.

There are several possible methods to get existing voices to speak in noise. One novel approach is to use *style conversion*. Using techniques that were designed for voice conversion between a source and target speaker [8], we applied such techniques to learn a mapping between plain speech and speech that was generated in noise. This work uses a Gaussian Mixture Model transformation method [9], and works primarily with the spectral differences between the two styles, as well as some minimal pitch and durational differences. It is important to reiterate that those differences are *not* all that distinguish speech in noise from plain speech, and thus the transformation model is not going to be able to produce natural-quality speech in noise.

Though a more parametric approach is likely to be better in the long run, the style conversion technique allows for relatively simple transformation to speech in noise. However, it does come at the cost of quality, although we feel it is still sufficiently good to evaluate the resulting synthetic speech.

3. EVALUATING MODIFIED VOICES

3.1. Experimental Setup and Implementation

We have performed an experiment designed to evaluate the effect of speech in noise on understandability. Participants were asked to listen to and transcribe recorded sentences over the telephone, under various conditions. Those conditions were: natural (human-produced) plain speech and speech in noise, and synthetic plain speech and speech modified through style conversion to be more like speech in noise. The synthetic speech was produced by a

The next 61B leaves Forbes and Murray at 3:20 pm.
There is a 28X leaving Fifth and Bellefield at 9:45 am.

Fig. 1. Example sentences from this evaluation showing the two different patterns.

limited-domain unit selection synthesizer designed specifically for the domain used in this evaluation, Pittsburgh bus information. All of the speech examples were power normalized to ensure that any differences we found were not due to volume. As a further condition, noise either was or was not added to the recordings; with added noise, the resulting signal-to-noise ratio is -3.2 dB. This gives a total of eight conditions; each subject transcribed one sentence from each of the conditions. Though all subjects heard the same eight sentences, they did not all hear them in the same order, nor did they have the same order of conditions. Two different sentence orders and four different condition orders were used. For every subject, however, all of the odd-numbered sentences had no noise added, and all of the even-numbered ones did. This provides eight different experiment "sequences", which were assigned to subjects based on their randomly assigned experiment number.

After transcribing the eight sentences, the participants were asked to complete a short questionnaire to provide information such as general age range, familiarity with the domain the sentences' content was from, whether they were a native speaker, and whether they had any hearing difficulties. Participants who completed the experiment were compensated with US$5.

The sentences in this study were from the domain of bus information, providing believable times with valid bus number / bus stop combinations for Pittsburgh's bus system. Two example sentences are shown in Figure 1. All of the sentences in the study have the pattern of one of the examples, changing the bus number, bus stop, and time. This domain is finite, but quite large when considering the bus route and stop coverage of the Port Authority. For this study, the sentences did not cover the full domain, but only a small fraction of it, using only 7 bus routes and 8 bus stops. However, participants were not aware of these limitations, and so any uncertainty would mean that people would have to consider the entire domain (or at least as much of it as they know) to disambiguate routes or stops.

To implement this study, we wrote a simple, single-file VoiceXML application. This file, along with our speech recordings, was then placed on the Internet. We then used a free commercial developer system to make our application available over the telephone. By calling a toll-free phone number, this commercial system would load our application via http and then execute it, allowing us to run the study.

There were some drawbacks to this implementation, however. First, though the commercial system worked flawlessly while we ran the experiment, there was the concern that we were dependent on an outside system that could stop working at any time. While this did not happen, it was still not an ideal situation. Additionally, because the VoiceXML server was not under our control, this limited our debugging ability during development, as well as limiting logging capability as the application was running. Though these were not significant problems, they did require some compromises in the design of the study and increased the development time. Furthermore, since this was a freely available developer system, the phone access was shared between many users. Though there were no busy signals, the shared access meant that accessing our specific application required first navigating through a few menus. This influenced the design of the study somewhat, because it meant that the partipants could not simply dial a phone number and do the task – the experimenter needed to go through the initial menus first.

3.2. Participant Groups

For this evaluation, we wanted to examine synthesis understandability in the general public, as this is one of the larger issues we have encountered. Furthermore, we wanted to examine how well even some extreme subgroups of the general public, such as elderly listeners, were able to understand speech synthesis, given the average age of the local population and the likely users of a bus information spoken dialog system. Elderly listeners present a different set of challenges to speech understandability than a typical evaluation group, such as graduate students. To that end, participants were divided into two groups: elderly and non-elderly, with the former defined as anyone age 60 or older, and the latter as anyone younger than 60. There were a total of 87 participants in this study, 45 of which were elderly and 42 of which were non-elderly. The non-elderly group, due to its similarity to a typical speech system evaluation group, is the baseline group for this study.

There are several things to note about these population groups. First, all of the elderly participants in this study came from Pittsburgh's senior citizen centers, which are buildings located in various city neighborhoods that elderly residents can go to during the day for social activites and events. This means that the elderly participants are moderately active, able to get around on their own, and in generally good health. As a group, they are a fairly accurate representation of the active elderly population in Pittsburgh. The non-elderly group primarily consists of university undergraduate and graduate students, as well as university staff, who answered a web-based solicitation for participants. Because of this, a significant percentage of non-elderly participants (approximately two-fifths) were not native speakers of English, which could negatively impact their performance on this task. Due to the methods of obtaining particpants for this study, the elderly group was predominately in their 60s and 70s, with a significant number of people in their 80s as well as a few in their 90s, while the young group is mostly people in their 20s and 30s, with a few in their 40s. The exact age statistics, as well as other demographic information, are shown in Table 1.

Non-Elderly		**Elderly**	
Participants	42	Participants	45
Age 18-29	27	Age 60-69	10
Age 30-39	11	Age 70-79	24
Age 40-49	4	Age 80+	11
Non-natives	16	Hearing Difficulties	15
Bus Riders	37	Bus Riders	37

Table 1. Age and other demographic information for the participants in this study.

We are assuming that, in general, the younger group has significantly fewer hearing problems than the elderly group. However, none of the participants, including the elderly, were given audiograms, due our need to travel to numerous locations within the city of Pittsburgh and the lack of access to a reasonably portable testing apparatus. We feel this assumption is justified when looking

at the young group as a whole, as none of the young participants reported any difficulties with their hearing. In contrast, one third of the elderly group self-reported hearing problems.

We did not make a distinction between which bus(es) a person rode when asking if they used the buses. Because the content of the sentences in this study dealt with buses and locations in the neighborhoods near Oakland (the neighborhood with several of Pittsburgh's universities), there is a concern that the non-elderly group would be more familiar with the specific bus numbers and stops used in the sentences, and be more likely to guess correctly when they had difficulty understanding what they heard. However, the elderly participants, despite not living in Oakland, have lived in the Pittsburgh area for many years (as compared to the non-elderly participants, who are overwhelmingly recent arrivals), and so the stop names should not be completely foreign to them. We did, however, track how often participants made use of the buses in general, with the majority riding the buses several times weekly, if not daily, though a significant minority (12% of the non-elderly and 17% of the elderly) did not ride the buses at all.

3.3. Results

We initially felt that Word Error Rate (WER) would be an appropriate measure to evaluate understandability. However, once people actually began to participate in this study, we discovered some issues that made us reconsider. First, a significant number of participants, especially those in the elderly group, did not follow the directions they were given; they did not write down every word they understood. Instead, these people wrote down only the bus number and bus stop, for example, despite being able to identify other words in the recordings they heard. This, obviously, has a negative effect on their WER scores, but those poor scores do not accurately reflect the understandability of the sentences, because words such as "the" and "is" are unnecessary to understand the meaning of the sentences. Because of that, it is not entirely clear that WER is measuring the right thing, and perhaps Concept Error Rate (CER) would be a more appropriate evaluation measure. We calculated both measures, and discovered that WER correlated well with the CER scores; thus, we feel comfortable reporting the understandability as measured by WER.

In each case, 'No Noise Added' means the original recordings were played, while 'Noise Added' means noise was added to the recordings such that the resulting signal-to-noise ratio was -3.2 dB. Additionally, 'Nat' refers to natural speech recordings, 'Syn' refers to synthetic speech examples, (P) refers to plain speech, and (N) refers to speech in noise.

Figures 2 and 3 show the overall WER scores for both populations for all 8 conditions. It is clear that the non-elderly group performs better on this task, with an approximately 20-40% better absolute WER than the elderly group. Further, these results show that natural speech in noise is more understandable than the plain speech, for both groups, though this result is not significant in all conditions. The synthetic speech, however, shows no significant increase in understandability in any condition, and shows significant *decrease* in understandability in three of the four conditions.

Noting the general difficulty the elderly group had with this task compared to the baseline, we attempted to isolate possible causes. Since the ability to hear the sentences is crucial to being able to understand them, and the elderly tend to have more hearing problems than the baseline group, we separated the elderly group based on whether they reported having hearing difficulties.

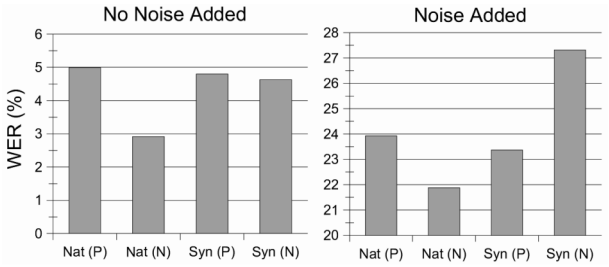

Fig. 2. Overall word error rate results for the baseline group.

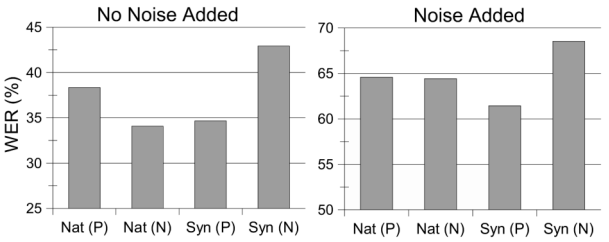

Fig. 3. Overall word error rate results for the elderly group.

It should be noted that this is potentially inaccurate, as some people with hearing problems either do not know, or will not admit, that they have them, and so subjects in the "normal hearing" group might have hearing deficiencies. Without performing a hearing test, which we did not do for this evaluation, there is little we can do besides trust the subjects to accurately describe their hearing. Figure 4 shows the results of these subgroups for the elderly population. It is clear from these results that people who report hearing problems perform significantly worse than those who claim to have no hearing problems.

Additionally, because the baseline group included a significant proportion of non-native listeners, we wanted to see what effect this had on the performance of the group. Figure 5 shows the results of the baseline group, separated into subgroups based on whether the subject's native language was English. Two things are immediately clear from this. First, as expected, native listeners are better at the task than non-natives. Secondly, speech in noise, especially natural speech in noise, shows WER reductions for native listeners, but does not with non-natives, even with naturally-produced speech.

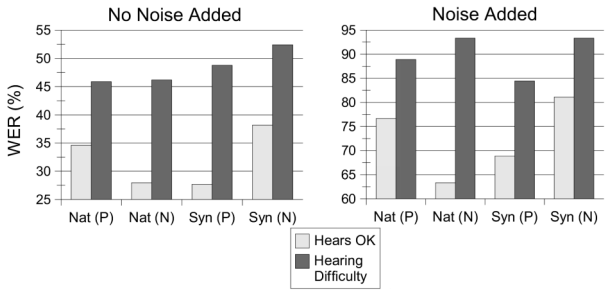

Fig. 4. Word error rate results for the elderly group, separated by self-reported hearing difficulties.

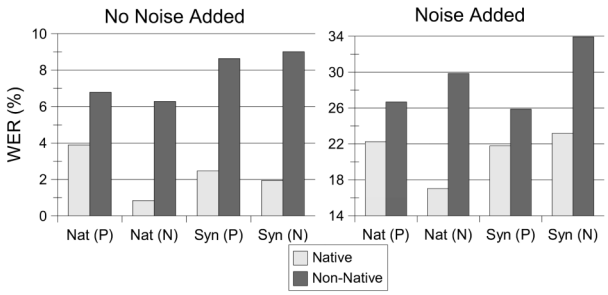

Fig. 5. Word error rate results for the baseline group, separated by nativeness.

We had also wanted to see if having knowledge of the domain, or at least if riding the buses in Pittsburgh offered any benefit. However, we did not have enough non-bus-riding participants to do any meaningful evaluation of this distinction.

4. DISCUSSION

There are several things to note about our results. First, natural speech in noise gives an understandability improvement in most of the conditions, as would be expected from previous work. The cases where it does not are the most challenging extremes: elderly people with hearing problems, and non-native listeners with added noise. For the elderly, it could be argued that, especially with the added noise, the word error rates are so high that the subjects were essentially guessing, if they managed to write something down at all. This view is supported by the fact that the increase in WER is only barely significant, and then only when noise was added to the speech. For non-natives, their ability to understand speech (natural or synthetic) in noisy conditions is lower, as one would expect with a clearly harder task.

It is disappointing to see that the style-converted synthetic speech was nearly universally harder to understand than the original plain speech, and quite often significantly less understandable. There are a number of possible explanations for this. First, as noted above, the style conversion we are doing to transform plain speech into speech in noise uses an incomplete model, and so does not capture all of the style differences present in speech in noise. That the resulting style is not the same as the naturally-produced speech in noise could reduce the understandability gains. Secondly, the conversion process introduces a noticeable quality degradation in the signal, due to the effect of the signal processing used 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 gained by the speech in noise modification is apparently lost by the signal processing or the incomplete model, or some combination of those factors.

However, the positive results from the natural speech in noise confirm that there are gains to be had from this sort of stylistic change. We have also determined that the increase in understandability is not solely due to the power differences of the speaking style. If the quality degredation of the style conversion can be reduced, and the model improved, we should see increases in understandability. If the model cannot be sufficiently improved, we may have to explore other methods, such as directly applying F_0, duration, and other models to the synthetic voice.

5. 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.

We would like to thank Mary Esther Van Shura of Pittsburgh Citiparks, Dave Milewski, director of the North Side Senior Community Center, Lynn Ford Adams, director of the Homewood Senior Community Center, and Jason Vastola, director of the Mt. Washington Senior Community Center, for their assistance in encouraging seniors to participate in our research.

6. REFERENCES

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

[2] B. Langner and A. Black, "An examination of speech in noise and its effect on understandability for natural and synthetic speech," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-LTI-04-187, 2004.

[3] 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.

[4] A. Raux, B. Langner, A. Black, and M. Eskenazi, "LET'S GO: Improving spoken dialog systems for the elderly and non-native," in *Eurospeech03*, Geneva, Switzerland, 2003.

[5] B. Langner and A. Black, "Improving the understandability of speech synthesis by modeling speech in noise," in *ICASSP05*, Philadelphia, PA, 2005.

[6] 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.

[7] 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.

[8] T. Toda, "High-quality and flexible speech synthesis with segment selection and voice conversion," Ph.D. dissertation, Nara Institute for Science and Technology, 2003.

[9] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," in *Eurospeech95*, Madrid, Spain, 1995, pp. 447–450.