An Examination of Speech In Noise and its Effect on Understandability for Natural and Synthetic Speech

Brian Langner and Alan W Black

CMU-LTI-04-187

FINAL

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Abstract

This report describes *speech in noise*, a speaking style employed by people to improve understandability when speaking in noisy conditions. Evidence of understandability improvements for natural speech is shown, including recent experimental data. As even the highest quality speech synthesizers can be difficult to understand, the viability of using this speaking style to improve understandability of synthetic speech is explored. Several techniques for obtaining and using speech in noise for speech synthesis are described, in addition to data and results from understandability tests.

Also provided is the CMU SIN database, a database of speech in noise recordings suitable for speech synthesis. The database is distributed with automatically segmented phonetic labels, obtained using the voice building scripts from the FestVox system. Furthermore, support for the Festival Speech Synthesis System is provided, including two pre-built voices (one that speaks in noise, and an otherwise identical plain speech voice) that can be used as-is.

Festival and FestVox are available at http://www.festvox.org/.

The CMU SIN speech corpus is available at http://www.festvox.org/cmu_sin/.

---

1. Introduction 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*, 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* [2], 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 at a rock concert will be different than speech produced with 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 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 describe *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, through techniques such as concatenative unit selection [3], 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, spoken dialog systems would become significantly more usable in non-research environments.

2. Speech In Noise for Speech Synthesis

2.1. Obtaining Speech In Noise

It would seem at first glance that recordings of speech in noise are relatively simple to obtain: just stick your voice talent in a room with the noise source you want and record their speech. 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 also having a noise source contaminating our recordings; what we require is recordings of *clean* speech in noise. In this report, 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, we designed a recording method [5] that would account for them.

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. In addition, for the work described in this report, we used a close-talking head-mounted microphone with the headphones, though other microphone types can certainly be used. The recording should be done in a quiet room, soundproof booth, or other environments normally used for recording a synthesis database. 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.

2.2. Building Voices that Speak in Noise

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 typical methods for F_0 extraction to give poor results, which in turn lowers the quality of the resulting synthesis.

Additionally, 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 unfeasible, but both speaking styles are required.

2.3. The CMU_SIN Speech Database

With the goal of creating a publicly accessible database of speech in noise suitable for synthesis research, we set out to design and record such a database. We chose to use a subset of the CMU ARCTIC [6] prompt set for our database – specifically the first 500 prompts of the "A" set – to provide a reasonably large, yet phonetically balanced data set while keeping the time required to record the database relatively low. Details of the size and units of this prompt set, relative to the CMU ARCTIC sets are shown in Table 1. The recording was done in a quiet room with a laptop and a head-mounted close-talking microphone with headphones.

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

Table 1. Number of various units in the CMU_SIN prompt set, compared to the CMU ARCTIC prompt sets.

Once the prompt set had been recorded, we used the FestVox voice building tools [7] to make two full unit selection synthesizers – one for plain speech, and one for speech in noise. CMU SphinxTrain [8] was used to build full HMM-based acoustic models from the recorded utterances for each database. Those speaker-specific models were used to perform forced alignment, allowing for automatic phonetic labeling. We did not perform any hand-correction of the automatic labels.

This work was publicly released in June 2004 as part of the CMU_SIN [5] speech corpus. The license under which this database is released is shown in Appendix A. We have not, at this time, released a "combined" plain speech/speech in noise voice, though if such a voice is desired, it is trivial to create one using the CMU_SIN data.

2.4. Applying Natural Speech In Noise to Synthetic Speech

While the technique outlined above is capable of producing high-quality synthetic speech in noise, it 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 [9]. 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 [10], 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 (GMM) transformation method [11], as distributed with the FestVox tools [7]. This method works primarily with the spectral differences between the two styles, as well as some minimal pitch and durational differences.

(Discuss modelling f0, duration, power, etc. of speech in noise, then applying those models to an existing voice)

3. Evaluating Understandability (I)

3.1. Evaluation Setup

In order to evaluate the effect of speech in noise on understandability, we performed an experiment designed to test understandability. Participants were asked to listen to recorded sentences over the telephone and transcribe them, under various conditions. Those conditions were: natural human-produced plain speech, natural speech in noise, synthetic plain speech, and synthetic speech modified with the style conversion technique described above to be more like speech in noise. The synthetic speech was produced by a 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 present in the recordings. This gives a total of eight conditions. Each subject was given eight different sentences to transcribe, one from each condition. 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 (modulo 8). The details of these sequences are shown in Table 2.

After completing 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. A copy of this questionnaire is shown in Appendix B. Participants who completed the experiment were compensated with $5.

The next 61B leaves Forbes and Murray at 3:20 pm.

There is a 28X leaving Fifth and Bellefield at 9:45 am.

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

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. The size of the domain, as well as the size of the subset used in the study, is shown in Table 3.

Sequence Number	*Sentence Order*	*Recordings Without Noise*	*Recordings With Noise*	*Sentence Conditions*
1	A	1, 3, 5, 7	2, 4, 6, 8	NAT-P, SYN-P, NAT-N, NAT-P, SYN-N, NAT-N, SYN-P, SYN-N
2	A	1, 3, 5, 7	2, 4, 6, 8	SYN-P, NAT-P, SYN-N, SYN-P, NAT-N, SYN-N, NAT-P, NAT-N
3	A	1, 3, 5, 7	2, 4, 6, 8	SYN-N, NAT-N, SYN-P, SYN-N, NAT-P, SYN-P, NAT-N, NAT-P
4	A	1, 3, 5, 7	2, 4, 6, 8	NAT-N, SYN-N, NAT-P, NAT-N, SYN-P, NAT-P, SYN-N, SYN-P
5	B	1, 3, 5, 7	2, 4, 6, 8	NAT-P, SYN-P, NAT-N, NAT-P, SYN-N, NAT-N, SYN-P, SYN-N
6	B	1, 3, 5, 7	2, 4, 6, 8	SYN-P, NAT-P, SYN-N, SYN-P, NAT-N, SYN-N, NAT-P, NAT-N
7	B	1, 3, 5, 7	2, 4, 6, 8	SYN-N, NAT-N, SYN-P, SYN-N, NAT-P, SYN-P, NAT-N, NAT-P
8	B	1, 3, 5, 7	2, 4, 6, 8	NAT-N, SYN-N, NAT-P, NAT-N, SYN-P, NAT-P, SYN-N, SYN-P

Table 2. Details for the experiment sequences given to subjects in this evaluation. "Recordings With Noise" refers to recordings which have had a noise source added. NAT-P is human-produced plain speech, NAT-N is human-produced speech in noise, SYN-P is the normal output of a unit selection synthesizer, and SYN-N is the same unit selection synthesizer modified towards speech in noise.

Domain Data Source	*# Bus Routes*	*# Bus Stops*
Full Port Authority Database	211	15002
Evaluation Subset	7	8

Table 3. Details of the size of the "bus information" domain, and the subset used in this evaluation.

3.2. Evaluation Implementation

To implement the study described above, we wrote a simple VoiceXML application; the file itself can be found in Appendix C. 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, via http, our application, and then execute it, allowing us to run the research study.

We chose VoiceXML over other possible solutions for a variety of reasons. First, it was quick and simple to write this implementation; including time to learn how to write VoiceXML, writing and debugging took under six hours. No other solution available to us at the time would have required less time than that. Secondly, though what we are doing here is essentially a very simple dialog system, we did not want to deal with speech recognition and its associated problems as a control mechanism. DTMF is a reliable and near-universally available input method, and the commercial system we used offered a trivial way to incorporate DTMF recognition into our application. Again, all of the other solutions we looked into would have required more time and effort to implement working DTMF controls.

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.

3.3. Participant Groups

For this evaluation, we wanted to examine synthesis understandability in the general public, as this is one of the issues we have encountered in the Let's Go project. 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; that is, we expect most people to be able to have similar performance as this group.

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 such as whether or not a participant rides any of the buses in Pittsburgh, are shown in Table 4.

Non-Elderly Group	
Participants	42
Age 18-29	27
Age 30-39	11
Age 40-49	4
Non-Natives	16
Bus Riders	37

Elderly Group	
Participants	45
Age 60-69	10
Age 70-79	24
Age 80+	11
Hearing Difficulties	15
Bus Riders	37

Table 4. Age and other demographic information for the participants in this study.

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. Table 5 shows the breakdown of how often the bus-riding participants make use of the bus.

Frequency of Bus Use	***Non-Elderly***	***Elderly***
Daily, one or more times	11	18
One to five times weekly	13	6
One to four times monthly	8	7
Less than once a month	5	6
TOTAL	37	37
No bus use at all	*5*	*8*

Table 5. Number of participants from each group that ride the bus at varying frequencies.

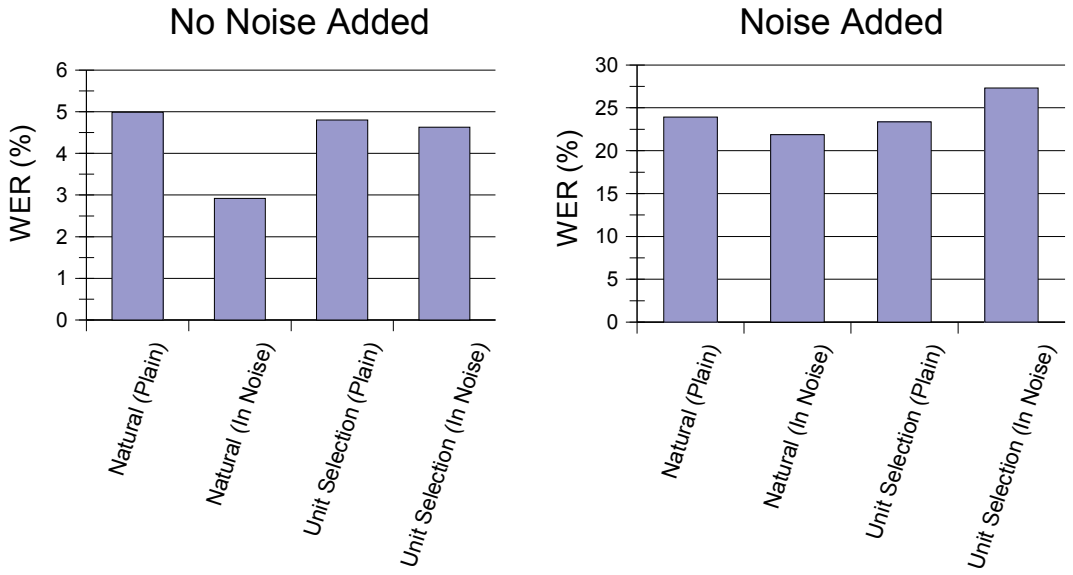
3.4. Results: Word Error Rate

We initially felt that Word Error Rate 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 what we want it to. Despite this, we still felt that word error rate could give at least a rough approximate of understandability.

3.4.1. Baseline (Non-Elderly) Scores

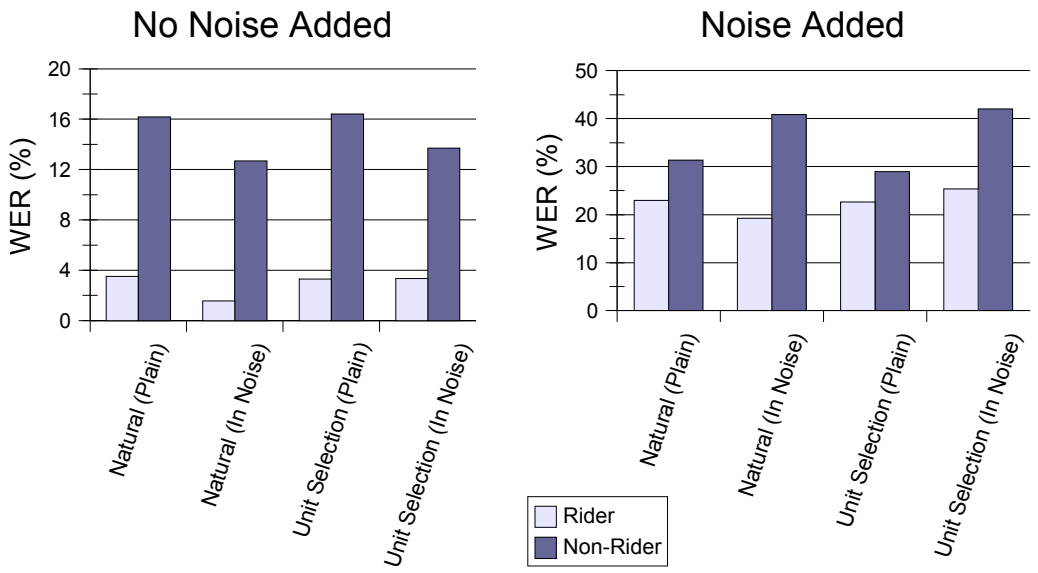
Included below are several figures and tables outlining the word error rate scores for the non-elderly group, both as a whole, and divided into a number of interesting subgroups, such as native versus non-native listeners and bus riders versus non-bus riders. In each case, 'No Noise Added' means the 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.



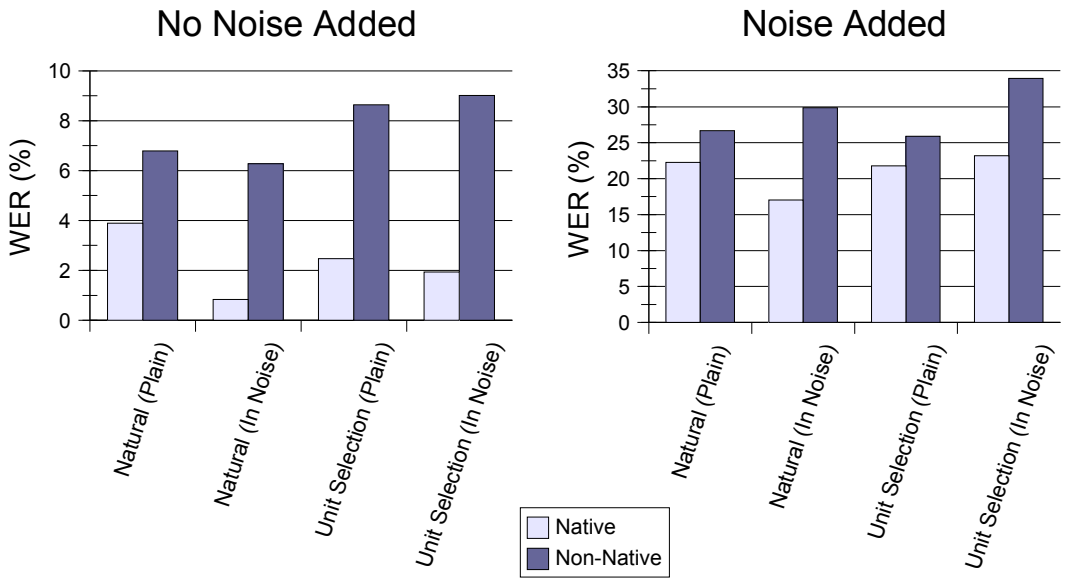
Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*WER (%)*
Natural	Plain	All Subjects	no noise	4.99
Natural	In Noise	All Subjects	no noise	2.92
Unit Selection	Plain	All Subjects	no noise	4.80
Unit Selection	In Noise	All Subjects	no noise	4.63
Natural	Plain	All Subjects	-3.2 dB S/N	23.93
Natural	In Noise	All Subjects	-3.2 dB S/N	21.88
Unit Selection	Plain	All Subjects	-3.2 dB S/N	23.37
Unit Selection	In Noise	All Subjects	-3.2 dB S/N	27.32

Figure 2. Graphs showing the average word error rate for each of the speech types and styles for the entire non-elderly group (42 subjects).



Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*WER (%)*
Natural	Plain	Riders	no noise	3.51
Natural	In Noise	Riders	no noise	1.57
Unit Selection	Plain	Riders	no noise	3.29
Unit Selection	In Noise	Riders	no noise	3.33
Natural	Plain	Non-Riders	no noise	16.18
Natural	In Noise	Non-Riders	no noise	12.68
Unit Selection	Plain	Non-Riders	no noise	16.42
Unit Selection	In Noise	Non-Riders	no noise	13.70
Natural	Plain	Riders	-3.2 dB S/N	22.97
Natural	In Noise	Riders	-3.2 dB S/N	19.26
Unit Selection	Plain	Riders	-3.2 dB S/N	22.61
Unit Selection	In Noise	Riders	-3.2 dB S/N	25.34
Natural	Plain	Non-Riders	-3.2 dB S/N	31.34
Natural	In Noise	Non-Riders	-3.2 dB S/N	40.85
Unit Selection	Plain	Non-Riders	-3.2 dB S/N	28.99
Unit Selection	In Noise	Non-Riders	-3.2 dB S/N	42.03

Figure 3. Graphs showing the effect of being a bus rider on word error rate for non-elderly participants for each of the speech type and style conditions in the evaluation. Note that there are only *5* non-riders (and 37 riders).

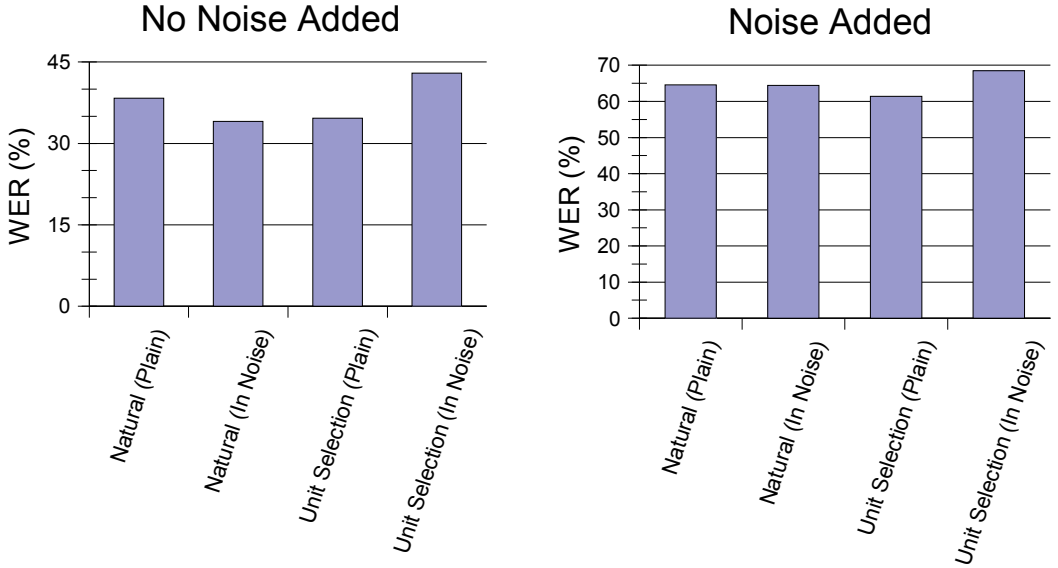


Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*WER (%)*
Natural	Plain	Natives	no noise	3.89
Natural	In Noise	Natives	no noise	0.84
Unit Selection	Plain	Natives	no noise	2.47
Unit Selection	In Noise	Natives	no noise	1.94
Natural	Plain	Non-Natives	no noise	6.79
Natural	In Noise	Non-Natives	no noise	6.28
Unit Selection	Plain	Non-Natives	no noise	8.64
Unit Selection	In Noise	Non-Natives	no noise	9.01
Natural	Plain	Natives	-3.2 dB S/N	22.25
Natural	In Noise	Natives	-3.2 dB S/N	17.03
Unit Selection	Plain	Natives	-3.2 dB S/N	21.79
Unit Selection	In Noise	Natives	-3.2 dB S/N	23.18
Natural	Plain	Non-Natives	-3.2 dB S/N	26.69
Natural	In Noise	Non-Natives	-3.2 dB S/N	29.86
Unit Selection	Plain	Non-Natives	-3.2 dB S/N	25.89
Unit Selection	In Noise	Non-Natives	-3.2 dB S/N	33.93

Figure 4. Graphs showing the effect of nativeness on word error rate for non-elderly participants for each of the speech type and style conditions in the evaluation. There are 26 natives and 16 non-natives.

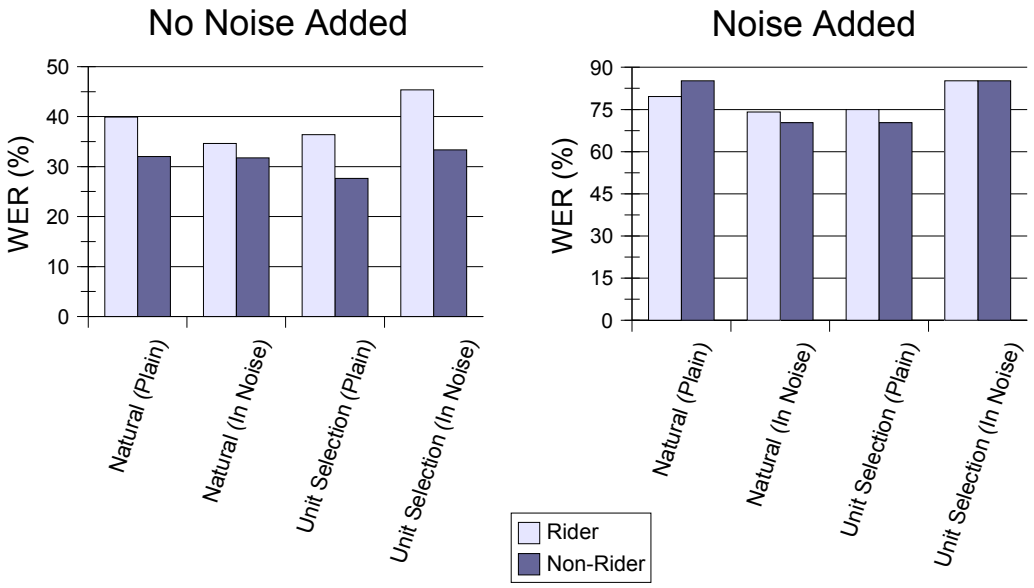
3.4.2. Elderly Scores

Included below are several figures and tables outlining the word error rate scores for the elderly group, both as a whole, and divided into a number of interesting subgroups, such as listeners who self-reported hearing difficulties versus those who did not and bus riders versus non-bus riders. Again, as with the non-elderly results, '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.



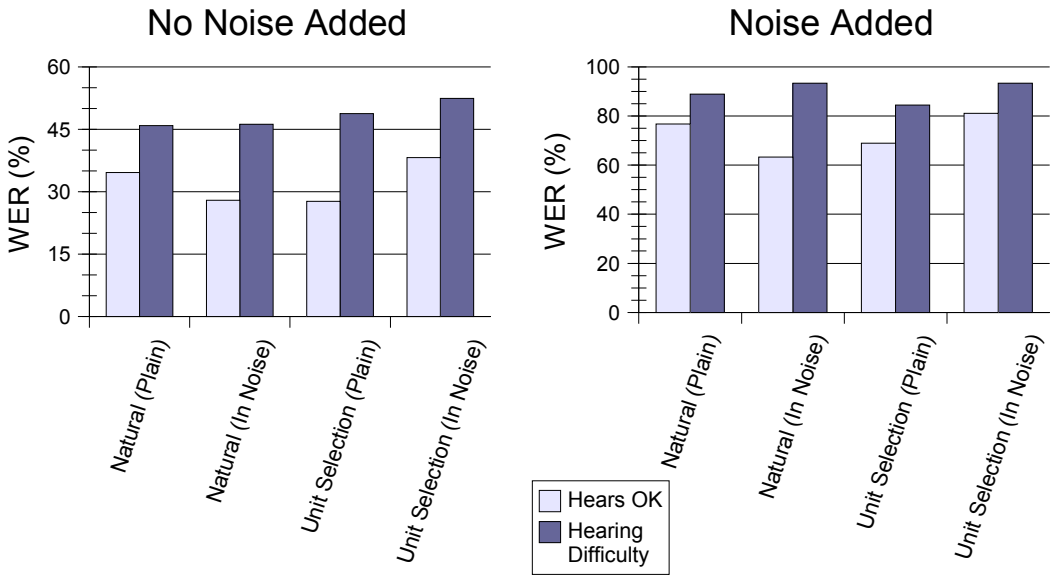
Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*WER (%)*
Natural	Plain	All Subjects	no noise	38.34
Natural	In Noise	All Subjects	no noise	34.08
Unit Selection	Plain	All Subjects	no noise	34.66
Unit Selection	In Noise	All Subjects	no noise	42.93
Natural	Plain	All Subjects	-3.2 dB S/N	64.58
Natural	In Noise	All Subjects	-3.2 dB S/N	64.41
Unit Selection	Plain	All Subjects	-3.2 dB S/N	61.44
Unit Selection	In Noise	All Subjects	-3.2 dB S/N	68.53

Figure 5. Graphs showing the average word error rate for each of the speech types and styles for the entire elderly group (45 subjects).



Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*WER (%)*
Natural	Plain	Riders	no noise	39.92
Natural	In Noise	Riders	no noise	34.67
Unit Selection	Plain	Riders	no noise	36.38
Unit Selection	In Noise	Riders	no noise	45.36
Natural	Plain	Non-Riders	no noise	32.00
Natural	In Noise	Non-Riders	no noise	31.75
Unit Selection	Plain	Non-Riders	no noise	27.64
Unit Selection	In Noise	Non-Riders	no noise	33.33
Natural	Plain	Riders	-3.2 dB S/N	64.07
Natural	In Noise	Riders	-3.2 dB S/N	65.93
Unit Selection	Plain	Riders	-3.2 dB S/N	62.60
Unit Selection	In Noise	Riders	-3.2 dB S/N	69.60
Natural	Plain	Non-Riders	-3.2 dB S/N	66.67
Natural	In Noise	Non-Riders	-3.2 dB S/N	58.40
Unit Selection	Plain	Non-Riders	-3.2 dB S/N	56.80
Unit Selection	In Noise	Non-Riders	-3.2 dB S/N	64.29

Figure 6. Graphs showing the effect of being a bus rider on word error rate for elderly participants for each of the speech type and style conditions in the evaluation. Note that there are only *8* non-riders (and 37 riders).



Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*WER (%)*
Natural	Plain	No Hearing Difficulties	no noise	34.61
Natural	In Noise	No Hearing Difficulties	no noise	27.95
Unit Selection	Plain	No Hearing Difficulties	no noise	27.68
Unit Selection	In Noise	No Hearing Difficulties	no noise	38.16
Natural	Plain	Hearing Difficulties	no noise	45.89
Natural	In Noise	Hearing Difficulties	no noise	46.19
Unit Selection	Plain	Hearing Difficulties	no noise	48.79
Unit Selection	In Noise	Hearing Difficulties	no noise	52.40
Natural	Plain	No Hearing Difficulties	-3.2 dB S/N	56.90
Natural	In Noise	No Hearing Difficulties	-3.2 dB S/N	59.37
Unit Selection	Plain	No Hearing Difficulties	-3.2 dB S/N	55.31
Unit Selection	In Noise	No Hearing Difficulties	-3.2 dB S/N	60.29
Natural	Plain	Hearing Difficulties	-3.2 dB S/N	80.39
Natural	In Noise	Hearing Difficulties	-3.2 dB S/N	74.29
Unit Selection	Plain	Hearing Difficulties	-3.2 dB S/N	73.46
Unit Selection	In Noise	Hearing Difficulties	-3.2 dB S/N	85.10

Figure 7. Graphs showing the effect of self-reported hearing difficulties on word error rate for elderly participants for each of the speech type and style conditions in the evaluation. There are 15 subjects with hearing difficulties and 30 without.

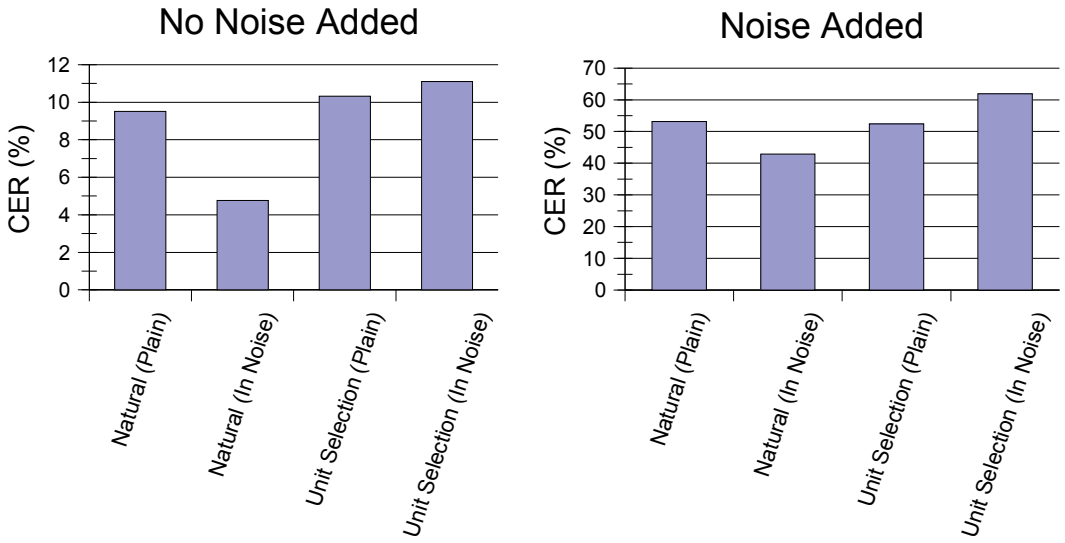
3.5. Results: Concept Error Rate

As discussed briefly above, it is not clear that Word Error Rate is actually measuring understandability. For speech recognition, WER is an ideal metric, because some speech recognition applications, such as dictation, require all words to be properly identified. There is, however, no such requirement for human-human speech interaction, and indeed, people can completely understand the meaning of what someone has said without understanding all of the words. This can be partly explained by the fact that natural language often contains "filler" words that are not strictly important to understanding the meaning of an utterance. The other words in the utterance are "content" words, and they convey the important concepts of the utterance. For a person to understand an utterance, only the content words need to be understood. For example, "predicting six inches snow tonight" can be easily understood despite being ungrammatical English, and conveys the same information as "The weather forecast is predicting up to six inches of snow to fall overnight tonight." The second utterance can be fully understood even with a WER of 67%, because only five of the fifteen words are conveying essential information. Worse, that also means that all of the important information could be lost with a WER of a little as 33%. In other words, the WER score is not useful in and of itself in determining if a person actually understood an utterance.

It is also apparent that people do not simply memorize the exact words that they hear. When asked to repeat what they have heard another person say, it is rare for a person to use the exact sentences and phrasing that they heard originally. Instead, people will describe the same concepts using different words – conveying the same information – but not an exact repetition. This lends credence to the idea that understanding does not require hearing every word correctly, but only the content words.

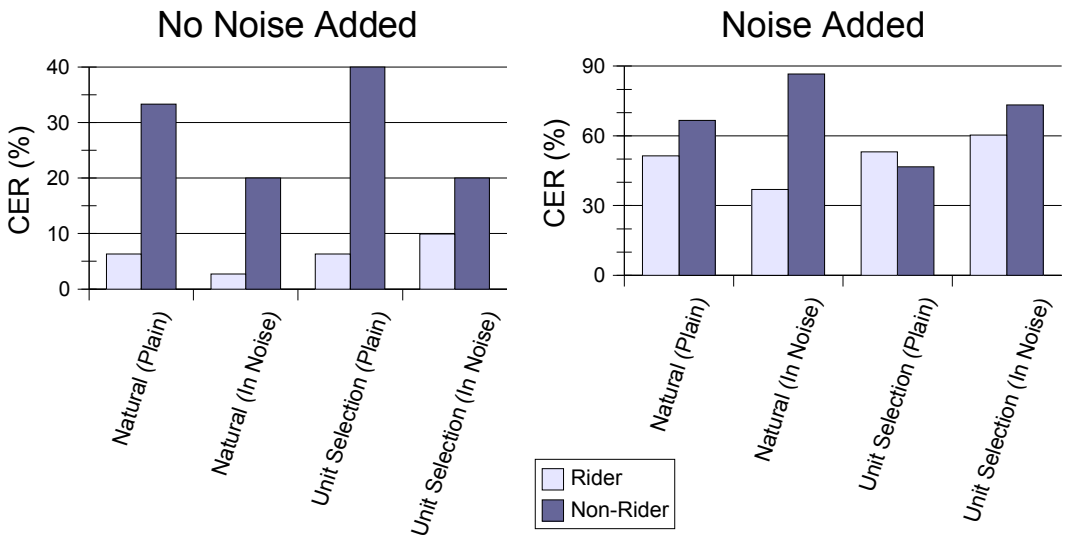
3.5.1. Baseline Scores

Included below are several figures and tables outlining the concept error rate scores for the non-elderly group, both as a whole, and divided into a number of interesting subgroups, such as native versus non-native listeners and bus riders versus non-bus riders. 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.



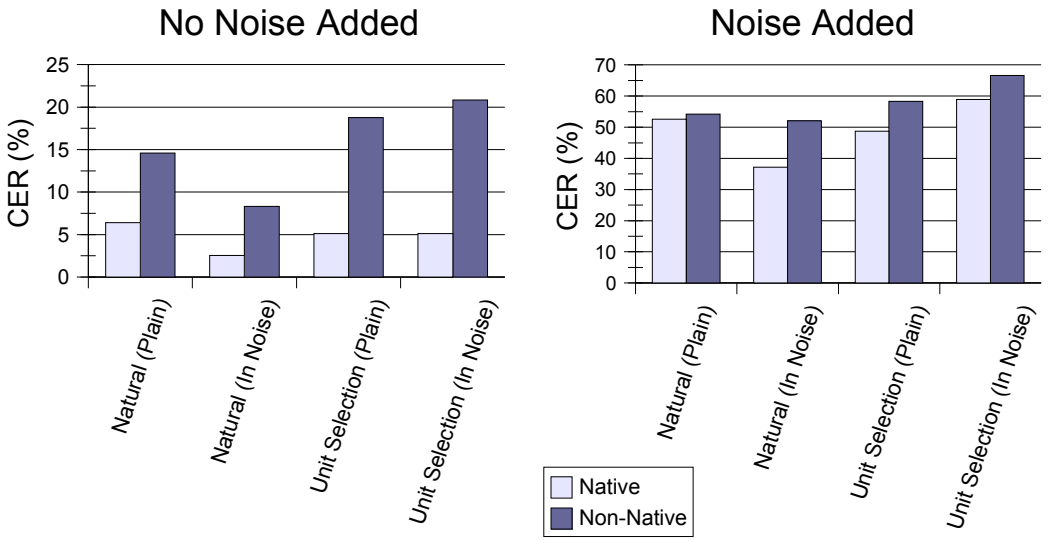
Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*CER (%)*
Natural	Plain	All Subjects	no noise	9.52
Natural	In Noise	All Subjects	no noise	4.76
Unit Selection	Plain	All Subjects	no noise	10.32
Unit Selection	In Noise	All Subjects	no noise	11.11
Natural	Plain	All Subjects	-3.2 dB S/N	53.17
Natural	In Noise	All Subjects	-3.2 dB S/N	42.86
Unit Selection	Plain	All Subjects	-3.2 dB S/N	52.38
Unit Selection	In Noise	All Subjects	-3.2 dB S/N	61.91

Figure 8. Graphs showing the average concept error rate for each of the speech types and styles for the entire non-elderly group (42 subjects).



Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*CER (%)*
Natural	Plain	Riders	no noise	6.31
Natural	In Noise	Riders	no noise	2.70
Unit Selection	Plain	Riders	no noise	6.31
Unit Selection	In Noise	Riders	no noise	9.91
Natural	Plain	Non-Riders	no noise	33.33
Natural	In Noise	Non-Riders	no noise	20.00
Unit Selection	Plain	Non-Riders	no noise	40.00
Unit Selection	In Noise	Non-Riders	no noise	20.00
Natural	Plain	Riders	-3.2 dB S/N	51.35
Natural	In Noise	Riders	-3.2 dB S/N	36.94
Unit Selection	Plain	Riders	-3.2 dB S/N	53.15
Unit Selection	In Noise	Riders	-3.2 dB S/N	60.36
Natural	Plain	Non-Riders	-3.2 dB S/N	66.67
Natural	In Noise	Non-Riders	-3.2 dB S/N	86.67
Unit Selection	Plain	Non-Riders	-3.2 dB S/N	46.67
Unit Selection	In Noise	Non-Riders	-3.2 dB S/N	73.33

Figure 9. Graphs showing the effect of being a bus rider on concept error rate for non-elderly participants for each of the speech type and style conditions in the evaluation. Note that there are only 5 non-riders (and 37 riders).

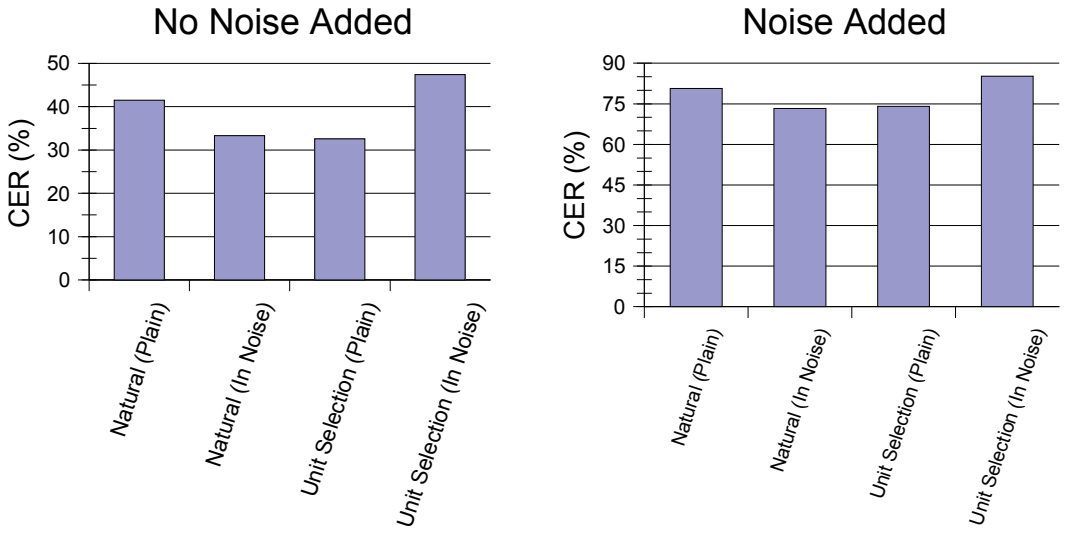


Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*CER (%)*
Natural	Plain	Natives	no noise	6.41
Natural	In Noise	Natives	no noise	2.56
Unit Selection	Plain	Natives	no noise	5.13
Unit Selection	In Noise	Natives	no noise	5.13
Natural	Plain	Non-Natives	no noise	14.58
Natural	In Noise	Non-Natives	no noise	8.33
Unit Selection	Plain	Non-Natives	no noise	18.75
Unit Selection	In Noise	Non-Natives	no noise	20.83
Natural	Plain	Natives	-3.2 dB S/N	52.56
Natural	In Noise	Natives	-3.2 dB S/N	37.18
Unit Selection	Plain	Natives	-3.2 dB S/N	48.72
Unit Selection	In Noise	Natives	-3.2 dB S/N	58.97
Natural	Plain	Non-Natives	-3.2 dB S/N	54.17
Natural	In Noise	Non-Natives	-3.2 dB S/N	52.08
Unit Selection	Plain	Non-Natives	-3.2 dB S/N	58.33
Unit Selection	In Noise	Non-Natives	-3.2 dB S/N	66.67

Figure 10. Graphs showing the effect of nativeness on concept error rate for non-elderly participants for each of the speech type and style conditions in the evaluation. There are 26 natives and 16 non-natives.

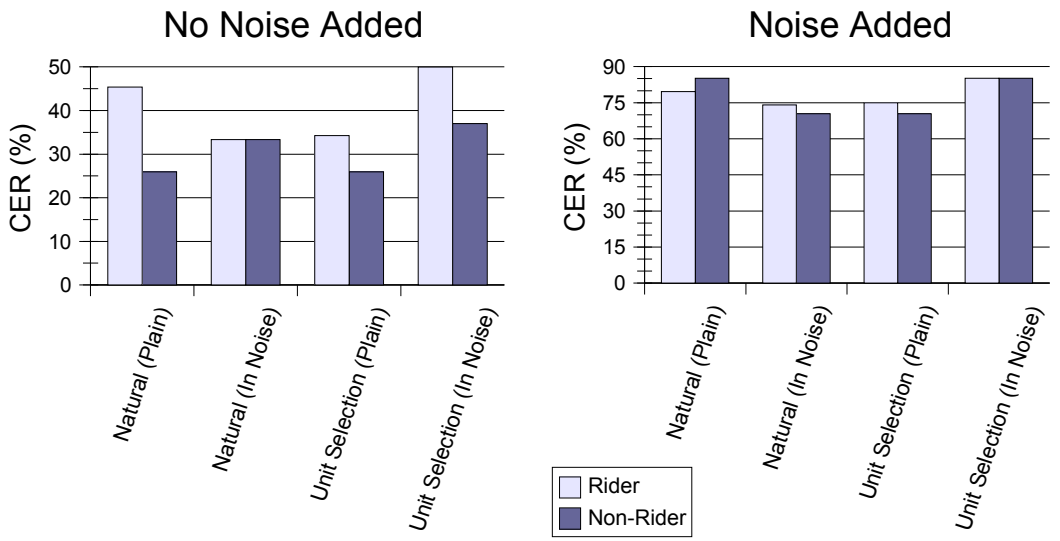
3.5.2. Elderly Scores

Included below are several figures and tables outlining the concept error rate scores for the elderly group, both as a whole, and divided into a number of interesting subgroups, such as listeners who self-reported hearing difficulties versus those who did not and bus riders versus non-bus riders. Again, as with the non-elderly results, '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.



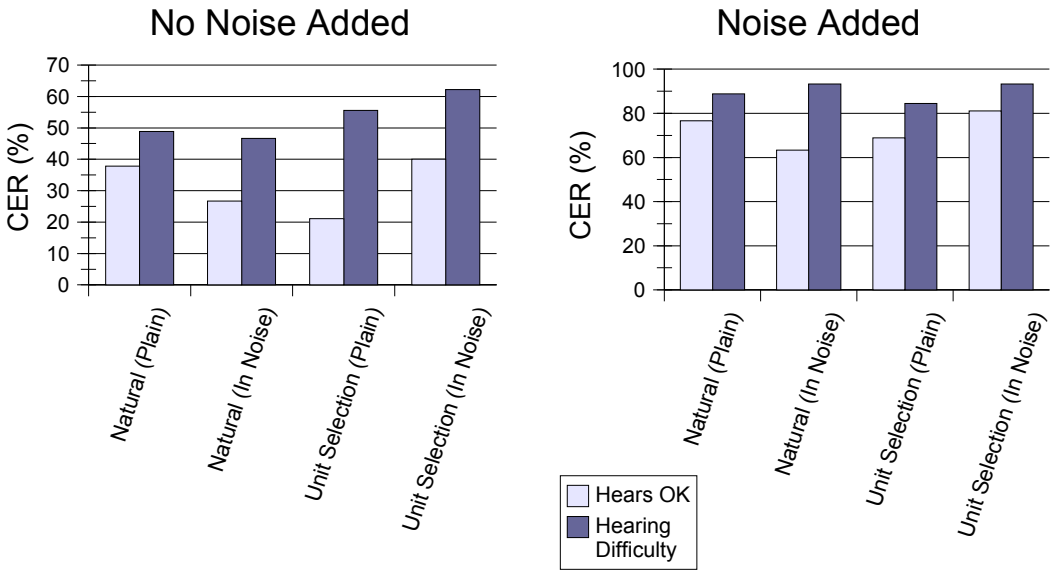
Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*CER (%)*
Natural	Plain	All Subjects	no noise	41.48
Natural	In Noise	All Subjects	no noise	33.33
Unit Selection	Plain	All Subjects	no noise	32.59
Unit Selection	In Noise	All Subjects	no noise	47.41
Natural	Plain	All Subjects	-3.2 dB S/N	80.74
Natural	In Noise	All Subjects	-3.2 dB S/N	73.33
Unit Selection	Plain	All Subjects	-3.2 dB S/N	74.07
Unit Selection	In Noise	All Subjects	-3.2 dB S/N	85.19

Figure 11. Graphs showing the average concept error rate for each of the speech types and styles for the entire elderly group (45 subjects).



Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*CER (%)*
Natural	Plain	Riders	no noise	45.37
Natural	In Noise	Riders	no noise	33.33
Unit Selection	Plain	Riders	no noise	34.26
Unit Selection	In Noise	Riders	no noise	50.00
Natural	Plain	Non-Riders	no noise	25.93
Natural	In Noise	Non-Riders	no noise	33.33
Unit Selection	Plain	Non-Riders	no noise	25.93
Unit Selection	In Noise	Non-Riders	no noise	37.04
Natural	Plain	Riders	-3.2 dB S/N	79.63
Natural	In Noise	Riders	-3.2 dB S/N	74.07
Unit Selection	Plain	Riders	-3.2 dB S/N	75.00
Unit Selection	In Noise	Riders	-3.2 dB S/N	85.19
Natural	Plain	Non-Riders	-3.2 dB S/N	85.19
Natural	In Noise	Non-Riders	-3.2 dB S/N	70.37
Unit Selection	Plain	Non-Riders	-3.2 dB S/N	70.37
Unit Selection	In Noise	Non-Riders	-3.2 dB S/N	85.19

Figure 12. Graphs showing the effect of being a bus rider on concept error rate for elderly participants for each of the speech type and style conditions in the evaluation. Note that there are only *8* non-riders (and 37 riders).



Speech Type	*Speech Style*	*Subject Group*	*Noise Level*	*CER (%)*
Natural	Plain	No Hearing Difficulties	no noise	37.78
Natural	In Noise	No Hearing Difficulties	no noise	26.67
Unit Selection	Plain	No Hearing Difficulties	no noise	21.11
Unit Selection	In Noise	No Hearing Difficulties	no noise	40.00
Natural	Plain	Hearing Difficulties	no noise	48.89
Natural	In Noise	Hearing Difficulties	no noise	46.67
Unit Selection	Plain	Hearing Difficulties	no noise	55.56
Unit Selection	In Noise	Hearing Difficulties	no noise	62.22
Natural	Plain	No Hearing Difficulties	-3.2 dB S/N	76.67
Natural	In Noise	No Hearing Difficulties	-3.2 dB S/N	63.33
Unit Selection	Plain	No Hearing Difficulties	-3.2 dB S/N	68.89
Unit Selection	In Noise	No Hearing Difficulties	-3.2 dB S/N	81.11
Natural	Plain	Hearing Difficulties	-3.2 dB S/N	88.89
Natural	In Noise	Hearing Difficulties	-3.2 dB S/N	93.33
Unit Selection	Plain	Hearing Difficulties	-3.2 dB S/N	84.44
Unit Selection	In Noise	Hearing Difficulties	-3.2 dB S/N	93.33

Figure 13. Graphs showing the effect of self-reported hearing difficulties on concept error rate for elderly participants for each of the speech type and style conditions in the evaluation. There are 15 subjects with hearing difficulties and 30 without.

3.6. Analysis of Data / Conclusions

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 directly to the synthetic voice, rather than using an intermediate process like style conversion to get a synthetic voice that speaks in noise.

4. Evaluating Understandability (II)

4.1. Evaluation Setup and Details

In addition to the telephone-based evaluation described above, we also

performed a "listening test" evaluation [12]. For this test, we placed about 100 recordings on the Internet and had ten people experienced with speech synthesis technology listen to them. The test involved two different styles of evaluation: first, a transcription task, where listeners were told to listen to the recordings and type in what they heard; secondly, a *mean opinion score* task, where listeners were told to listen to the recordings and rate them on a scale of 1 to 5, 5 being best. Like the previous evaluation, there were several different conditions the recordings could fall into; in addition to the eight conditions used previously, this evaluation also used diphone-based synthetic speech. As with the unit-selection-based speech, the diphone speech was modified with the style conversion technique to speak in noise as well.

Furthermore, there were two levels of noise added to recordings, rather than just the one level in the previous evaluation. These noise levels were chosen based on the results of a small empirical study where people were asked to listen to recordings with added noise that had a wide range of signal-to-noise ratios (from -12.1 dB to +8.7 dB). Noise level 1 was point at which most people in the study could get some words correct while also making some mistakes, and noise level 2 was the level at which most people could not understand more than a couple of words. Noise level 1 corresponds to a -3.2 dB SNR, and noise level 2 corresponds to a -4.9 dB SNR. Note that because the power of the added noise is greater than that of the original speech, the signal-to-noise ratios are negative. These variables (3 types of speech, 2 speaking styles, 3 noise levels) provide for a total of 18 different conditions used in the evaluation. Table 6 outlines these 18 conditions in detail.

The sentences used in this evaluation are the same as those used in the evaluation described above; the sample sentences shown in Figure 1 are again representative of the information those sentences contain. The content of the recordings was identical in all cases: a single natural, plain speech sentence of bus information. The added noise is designed to make the task harder in order to more clearly identify any performance differences. 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. 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.

Condition	*Description*
NAT-P@0	Natural plain speech, no noise added
DIPH-P@0	Synthetic diphone-based plain speech, no noise added
UNIT-P@0	Synthetic unit-selection-based plain speech, no noise added
NAT-P@1	Natural plain speech, noise added to give SNR of -3.2 dB
DIPH-P@1	Synthetic diphone-based plain speech, noise added to give SNR of -3.2 dB
UNIT-P@1	Synthetic unit-selection-based plain speech, noise added to give SNR of -3.2 dB
NAT-P@2	Natural plain speech, noise added to give SNR of -4.9 dB
DIPH-P@2	Synthetic diphone-based plain speech, noise added to give SNR of -4.9 dB
UNIT-P@2	Synthetic unit-selection-based plain speech, noise added to give SNR of -4.9 dB
NAT-N@0	Natural speech in noise, no noise added
DIPH-N@0	Synthetic diphone-based speech in noise, no noise added
UNIT-N@0	Synthetic unit-selection-based speech in noise, no noise added
NAT-N@1	Natural speech in noise, noise added to give SNR of -3.2 dB
DIPH-N@1	Synthetic diphone-based speech in noise, noise added to give SNR of -3.2 dB
UNIT-N@1	Synthetic unit-selection-based speech in noise, noise added to give SNR of -3.2 dB
NAT-N@2	Natural speech in noise, noise added to give SNR of -4.9 dB
DIPH-N@2	Synthetic diphone-based speech in noise, noise added to give SNR of -4.9 dB
UNIT-N@2	Synthetic unit-selection-based speech in noise, noise added to give SNR of -4.9 dB

Table 6. Detailed description of the 18 conditions present in this evaluation.

4.2. Results: Word Error Rate

Listeners were asked to listen to four examples of each noise level / speaking style combination for naturally produced speech, for a total of 24 sentences. The sentences were arranged randomly, with the stipulation that the same condition could not be heard in two consecutive sentences. The listeners were asked to listen to each sentence as few times as possible. In nearly all cases, this was fewer than three times per sentence. Their answers were then scored using Word Error Rate (WER); the results are shown in Figure 14, below.

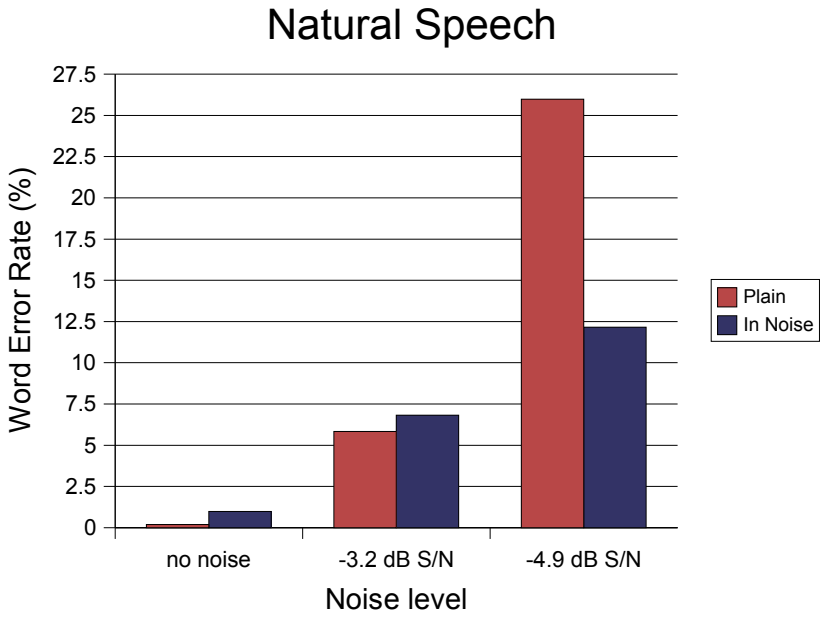


Figure 14. Word Error Rate scores for natural speech, in plain and in-noise styles, at various noise levels.

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 performed on the recordings. This suggests that the spectral, prosodic, and durational differences of speech in noise 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.

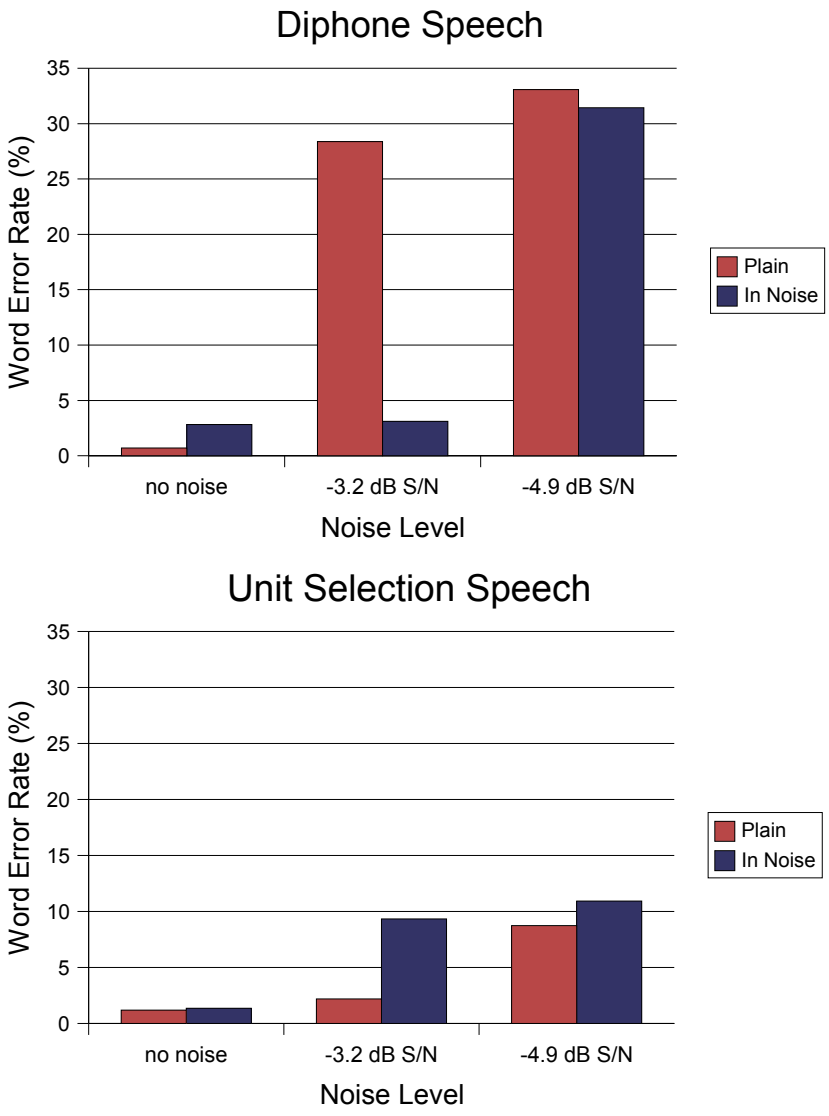
To evaluate the effectiveness of our modified synthetic speech, we used a similar process as with the natural speech in noise. Again, to account for power differences, all of the samples were power normalized to the level of natural plain

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 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 for the diphone and unit selection synthetic speech are shown below in Figure 15, respectively. Additionally, a graph showing all of the data, from both natural and synthetic speech, is in Figure 16.

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 possible 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.



Speaking Style	*S/N Ratio*	*Diphone WER (%)*	*Unit Selection WER (%)*
Plain Speech	no noise	0.70	1.19
Speech In Noise	no noise	2.82	1.36
Plain Speech	-3.2 dB	28.38	2.20
Speech In Noise	-3.2 dB	3.11	9.33
Plain Speech	-4.8 dB	33.07	8.73
Speech In Noise	-4.8 dB	31.43	10.92

Figure 15. Word Error Rate scores for diphone and unit selection synthetic speech, in plain and in-noise styles, at various noise levels.

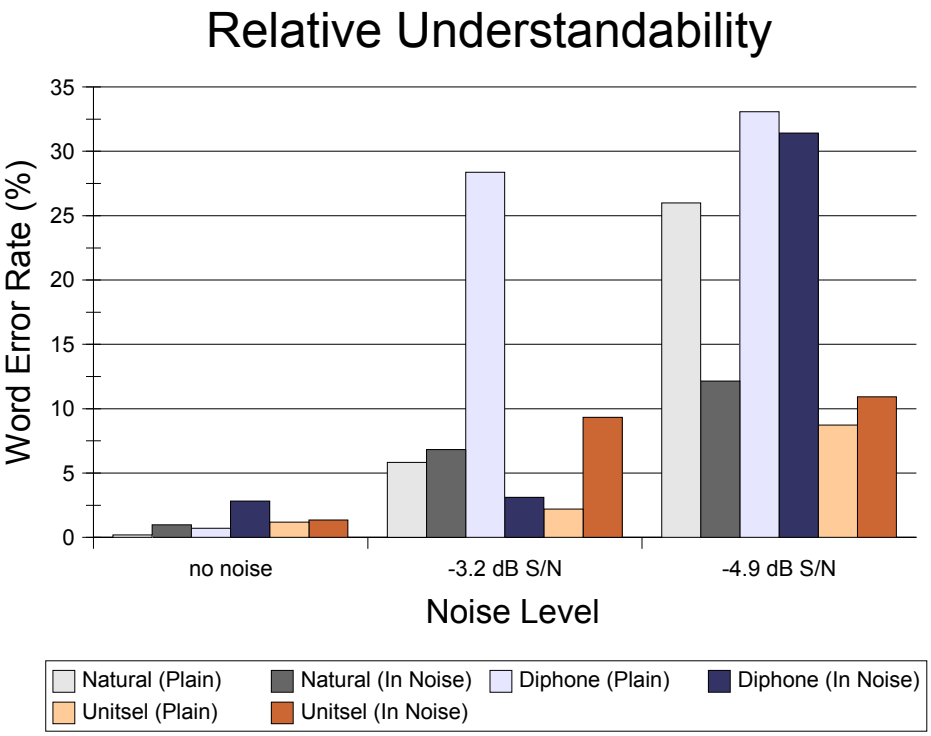


Figure 16. Word Error Rate scores for both natural and synthetic speech, in plain and in-noise styles, at various noise levels.

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.

However, these results also show that, despite the enormous improvement in the diphone voice, the unmodified unit selection voice still has a better word error rate. In fact, the plain-speech unit selection voice outperformed even natural speech (both plain and in noise) once there was noise added to the sentences.

4.3. Results: Concept Error Rate

As with the previous evaluation, Word Error Rate is not an ideal metric for measuring understandability. Though the listeners in this evaluation were speech

professionals familiar with this type of task and understood how their answers would be scored, there were still occasional cases where the listener did not transcribe all of what they had heard.

The same data evaluated in the word error rate section above was evaluated using concept error rate. Data entered by listeners was manually "collapsed" into concepts, and words that did not fit cleanly into a concept were simply removed. Each sentence had a total of 3 concepts: the bus number, the bus stop, and the time. There were no sentences entered by a listener that had more concepts than those three. The concept error rate scores for natural speech are shown below in Figure 17, and the results for the synthetic voices are shown in Figure 18. The exact scores are shown in Table 7. A graph of data from all three voice types is in Figure 19.

Speaking Style	*S/N Ratio*	*Natural CER (%)*	*Diphone CER (%)*	*Unit Selection CER (%)*
Plain Speech	no noise	0.88	2.22	3.33
Speech In Noise	no noise	2.63	10.00	1.23
Plain Speech	-3.2 dB	20.18	49.38	8.97
Speech In Noise	-3.2 dB	24.56	13.58	34.44
Plain Speech	-4.8 dB	44.74	65.43	61.73
Speech In Noise	-4.8 dB	35.96	56.67	27.78

Table 7. Concept Error Rate scores for natural speech, and diphone and unit selection synthetic speech, in plain and in-noise styles, at various noise levels.

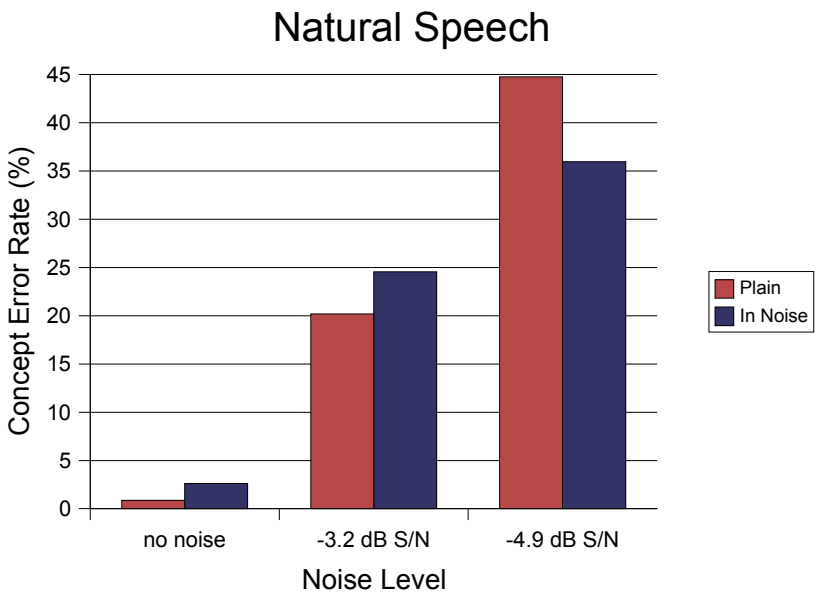
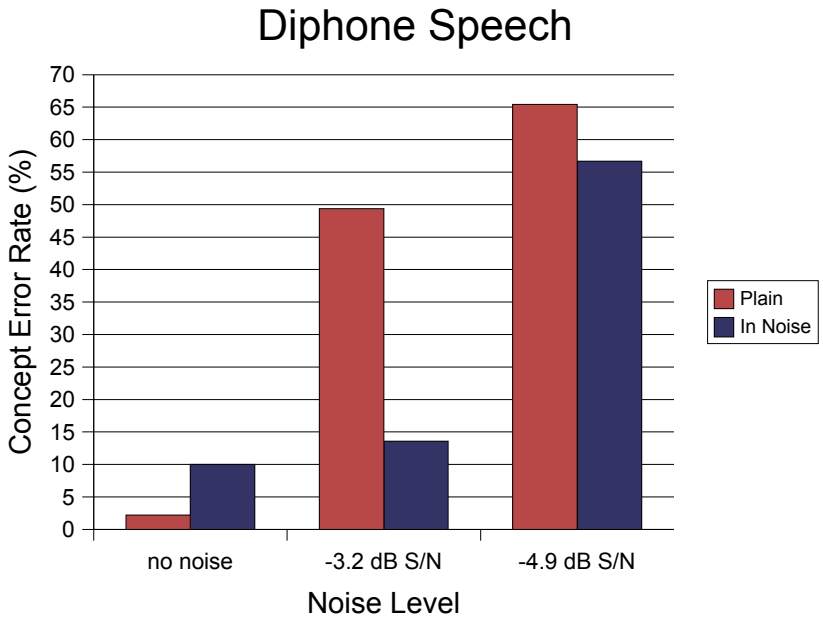


Figure 17. Concept Error Rate scores for natural speech, in plain and in-noise styles, at various noise levels.



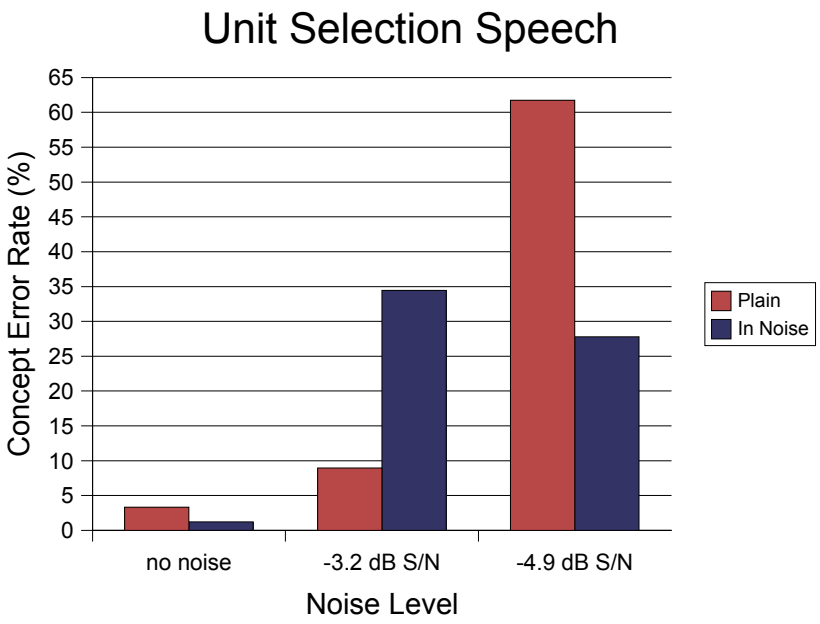


Figure 18. Concept Error Rate scores for synthetic unit-selection-based speech, in plain and in-noise styles, at various noise levels.

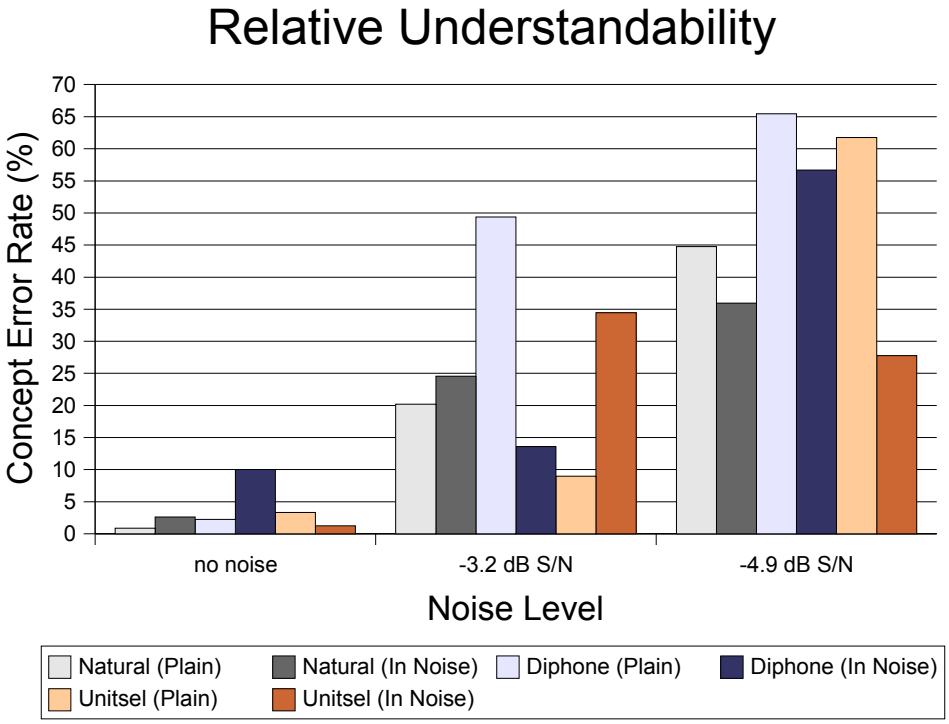


Figure 19. Concept Error Rate scores for both natural and synthetic speech, in plain and in-noise styles, at various noise levels.

4.4. Results: Mean Opinion Score

As a part of this evaluation, we also collected Mean Opinion Score data for each of the voice-style combinations. Listeners were asked to rate four examples of each combination (for a total of 24 sentences) on a scale of 1 to 5, 5 being best. No noise was added to these recordings, as was done in the other parts of this evaluation.

The motivation for collecting this data was twofold: to verify our intuitions about the quality level of the speech resulting from the style conversion, and to see if there is any correlation between speech people prefer to listen to and speech people can understand. A graph of the average score for each voice / style type is shown in Figure 20.

There are several things to note from this data. First, without any background noise, speech in noise is less preferred than plain speech, for each of the different voice types, including the naturally-produced speech. This is not particularly surprising, since the spectral and prosodic differences present in speech in noise are both noticeable and odd-sounding when heard under noiseless conditions. Further, the quality degradation in the unit selection voice is apparently quite perceptible, with a significant dropoff in score from the unmodified plain speech to the style-converted speech in noise. It should be pointed out that the diphone speech, though far less liked than the unit selection speech even when unmodified, does not show any significant reduction in likability when the style conversion is applied. This would seem to support the assertion that the inherent quality of the original diphone speech is not very different from the vocoder quality output from the style.

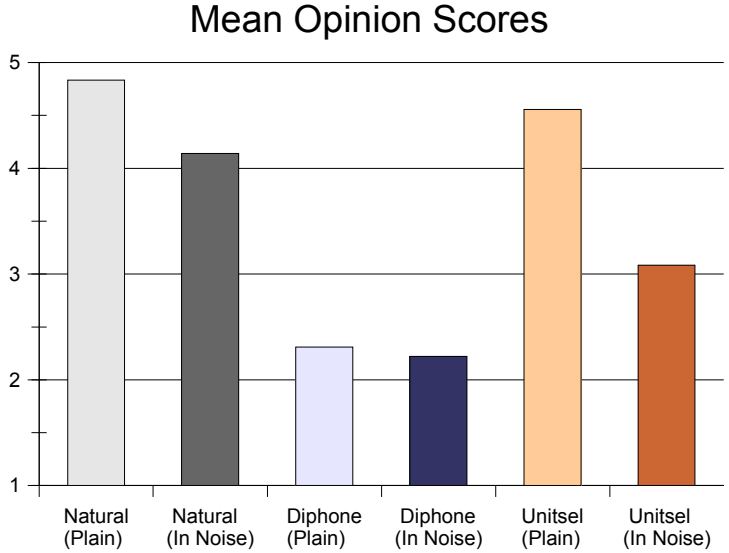


Figure 20. Mean opinion scores of the various voice / speaking style combinations used in this evaluation. Recordings were rated from 1 to 5, with 5 being best.

4.5. Analysis of Data / Conclusions

This evaluation 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.

Furthermore, 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.

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.

5. 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.

We would like to thank Maureen Bertocci and Megan Huff of the Port Authority of Allegheny County for their help in this work.

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 Mount Washington Senior Community Center for their assistance in encouraging seniors to participate in our research.

We would like to thank Stefanie Tomko for her contributions to the research studies described in this report.

Appendix A: Licensing

The CMU SIN database is distributed as "free software" under the following terms.

Appendix B: Questionnaire

The following is the questionnaire participants were asked to complete.

1. What is your age?
   _____ Under 18
   _____ 18-29
   _____ 30-39
   _____ 40-49
   _____ 50-59
   _____ 60-69
   _____ 70-79
   _____ 80 or older

2. Is English your first (native) language?
   _____ Yes _____ No

3. How familiar are you with synthetic (computer) voices?
   _____ I make them.
   _____ Very familiar.
   _____ Somewhat familiar.
   _____ Not very familiar.
   _____ I've heard them a couple times before.
   _____ I never heard one before today.

4. Do you have any hearing impairments or difficulties?
   _____ Yes _____ No

5. Do you ride public buses (or the T) in Pittsburgh?
   _____ Yes _____ No

6. If yes, about how often do you ride the bus?
   _____ Less than once a month
   _____ Between 1 and 4 times a month
   _____ Between 1 and 5 times a week
   _____ Once or twice a day
   _____ More than twice a day
   _____ Other (please explain) _____

7. What about this study did you think was easy? _____
_____

8. What about this study did you think was hard? _____
_____

Appendix C: experiment.vxml

The following is the VoiceXML file that is used to implement the experiment detailed in this report.

```
<vxml version="2.0">
<property name="inputmodes" value="dtmf"/>
<property name="bargein" value="false"/>
<property name="termtimeout" value="500ms"/>
<property name="timeout" value="10s"/>
<property name="universals" value="none"/>
<var name="expnum"/>
<script>
  var natpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/natural/";
  var natnpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/natural-noise/";
  var lgpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/letsgo/";
  var lgnpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/letsgo-noise/";
  var natNpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/naturalN/";
  var natNnpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/naturalN-
noise/";
  var lgNpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/letsgoN/";
  var lgNnpath =
"http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/letsgoN-noise/";
  var audiofilearray = [ natpath+"18.wav",
  natpath+"17.wav",
  natpath+"26.wav",
  natpath+"22.wav",
  natpath+"15.wav",
  natpath+"07.wav",
  natpath+"21.wav",
  natpath+"30.wav",

  natnpath+"18.wav",
  natnpath+"17.wav",
  natnpath+"26.wav",
  natnpath+"22.wav",
  natnpath+"15.wav",
  natnpath+"07.wav",
  natnpath+"21.wav",
  natnpath+"30.wav",

  lgpath+"18.wav",
  lgpath+"17.wav",
  lgpath+"26.wav",
  lgpath+"22.wav",
  lgpath+"15.wav",
  lgpath+"07.wav",
  lgpath+"21.wav",
  lgpath+"30.wav",
```

```
  lgnpath+"18.wav",
  lgnpath+"17.wav",
  lgnpath+"26.wav",
  lgnpath+"22.wav",
  lgnpath+"15.wav",
  lgnpath+"07.wav",
  lgnpath+"21.wav",
  lgnpath+"30.wav",

  natNpath+"18.wav",
  natNpath+"17.wav",
  natNpath+"26.wav",
  natNpath+"22.wav",
  natNpath+"15.wav",
  natNpath+"07.wav",
  natNpath+"21.wav",
  natNpath+"30.wav",


  natNnpath+"18.wav",
  natNnpath+"17.wav",
  natNnpath+"26.wav",
  natNnpath+"22.wav",
  natNnpath+"15.wav",
  natNnpath+"07.wav",
  natNnpath+"21.wav",
  natNnpath+"30.wav",

  lgNpath+"18.wav",
  lgNpath+"17.wav",
  lgNpath+"26.wav",
  lgNpath+"22.wav",
  lgNpath+"15.wav",
  lgNpath+"07.wav",
  lgNpath+"21.wav",
  lgNpath+"30.wav",

  lgNnpath+"18.wav",
  lgNnpath+"17.wav",
  lgNnpath+"26.wav",
  lgNnpath+"22.wav",
  lgNnpath+"15.wav",
  lgNnpath+"07.wav",
  lgNnpath+"21.wav",
  lgNnpath+"30.wav" ];

  var sequences = new Array();
  sequences.push(new Array(0, 25, 34, 11, 52, 45, 22, 63));
  sequences.push(new Array(16, 9, 50, 27, 36, 61, 6, 47));
  sequences.push(new Array(48, 41, 18, 59, 4, 29, 38, 15));
  sequences.push(new Array(32, 57, 2, 43, 20, 13, 54, 31));
  sequences.push(new Array(7, 24, 33, 10, 51, 44, 21, 62));
  sequences.push(new Array(23, 8, 49, 26, 35, 60, 5, 46));
  sequences.push(new Array(55, 40, 17, 58, 3, 28, 37, 14));
  sequences.push(new Array(39, 56, 1, 42, 19, 12, 53, 30));
</script>
```

```
<form id="start">
  <field name="expnum" modal="true" type="number">
    <prompt><audio
src="http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/stef4.wav">
Please enter your experiment number now</audio></prompt>
    <catch event="noinput nomatch">
      <audio
src="http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/stef5.wav">
Invalid experiment number, please try again.</audio>
      <reprompt/>
    </catch>
    <filled>
      <if cond="expnum &lt; 100 || expnum &gt; 500">
         <prompt><audio
src="http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/stef5.wav">
Invalid experiment number, try again.</audio></prompt>
       <clear namelist="expnum"/>
      <else/>
         <prompt><audio
src="http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/stef1.wav">
There is a problem with the
  recordings, please tell the experimenter.</audio></prompt>
       <prompt><audio
src="http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/stef2.wav">
</audio></prompt>
        <if cond="parseInt(dialog.expnum, 10) == 100">
         <assign name="document.expnum" expr="1"/>
        <else/>
         <assign name="document.expnum" expr="1+(dialog.expnum%8)"/>
        </if>
        <goto next="#experiment"/>
      </if>
    </filled>
  </field>
</form>

<form id="experiment">
<property name="timeout" value="180s"/>
<property name="termchar" value=" "/>
<property name="termtimeout" value="0s"/>
<var name="sentcount" expr="0"/>
<var name="rpeat" expr="0"/>

  <field name="advance" modal="true">
  <prompt> </prompt>
  <grammar mode="dtmf">
  <![CDATA[
  [
     [dtmf-1 dtmf-2 dtmf-3 dtmf-4 dtmf-5 dtmf-6 dtmf-7 dtmf-8 dtmf-9
  dtmf-0 dtmf-pound] {<advance "100">}
     [dtmf-star] {<advance "1">}
     [(dtmf-pound dtmf-7)] {<advance "2">}
     [(dtmf-pound dtmf-1)] {<advance "0">}
  ]
  ]]>
  </grammar>
  <catch event="noinput nomatch">
    <clear namelist="advance"/>
```

```
  </catch>

  <filled>
     <if cond="parseInt(advance, 10) == 0">
       <if cond="parseInt(rpeat,10) &gt;= 2">
         <prompt>No more repeats, Are allowed.</prompt>
       <clear namelist="advance"/>
       <else/>
         <assign name="sentcount" expr="sentcount-1"/>
         <prompt><audio expr="audiofilearray[sequences[expnum-1]
[sentcount]]">
           There is a problem with the recordings, please tell the
experimenter.
         </audio></prompt>
         <assign name="sentcount" expr="sentcount+1"/>
         <assign name="rpeat" expr="rpeat+1"/>
         <clear namelist="advance"/>
       </if>
     <elseif cond="parseInt(advance, 10) == 100"/>
       <prompt> </prompt>
       <clear namelist="advance"/>
     <elseif cond="parseInt(advance, 10) == 2 || parseInt(sentcount,
10) &gt;= 8"/>
         <goto next="#finish"/>
     <else/>
       <!-- play proper audio file or error message if there is a
problem -->
       <prompt><audio expr="audiofilearray[sequences[expnum-1]
[sentcount]]">
         There is a problem with the recordings, please tell the
experimenter.
       </audio></prompt>
       <assign name="sentcount" expr="sentcount+1"/>
       <assign name="rpeat" expr="0"/>
       <clear namelist="advance"/>
     </if>
  </filled>

  </field>
</form>

<form id="finish">
  <block>
     <audio
src="http://www.speech.cs.cmu.edu/letsgo/summer2004/wav/stef3.wav">
</audio>
     <disconnect/>
  </block>
</form>
</vxml>
```

References

[1] M. Eskenazi and A. Black, "A Study on Speech Over the Telephone and Aging," in *Eurospeech01*, Aalborg, Denmark, 2001.

[2] 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.

[3] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," in *ICASSP96*, Atlanta, GA, 1996.

[4] A. Raux, B. Langner, A. Black, and M. Eskenazi, "LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives," in *Eurospeech03*, Geneva, Switzerland, 2003.

[5] 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.

[6] J. Kominek and A. Black, "The CMU ARCTIC Speech Databases for Speech Synthesis Research," Tech. Report CMU-LTI-03-177 http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[7] A. Black and K. Lenzo, "Building Voices in the Festival Speech Synthesis System," http://festvox.org/bsv/, 2000.

[8] Carnegie Mellon University, "SphinxTrain: Building Acoustic Models for CMU Sphinx," http://www.speech.cs.cmu.edu/SphinxTrain/, 2001.

[9] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Pichney, and J. Pitrelli, "A Corpus-Based Approach to <AHEM/> Expressive Speech Synthesis," in *5th ISCA Workshop on Speech Synthesis*, Pittsburgh, PA, 2004.

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

[11] Y. Stylianou, O. Capp'e, and E. Moulines, "Statistical Methods for Voice Quality Transformation," in *Eurospeech95*, Madrid, Spain, 1995.

[12] B. Langner and A. Black, "Improving the Understandability of Speech Synthesis by Modeling Speech in Noise," in *ICASSP2005*, Philadelphia, PA, 2005.