

Improving Speech Synthesis for High Intelligibility under Adverse Conditions

Davis Pan, Brian Heng, Shiufun Cheung, and Ed Chang

Cambridge Research Laboratory

Compaq Computer Corporation, One Cambridge Center, Cambridge, MA 02142-1612

ABSTRACT

We investigate methods of improving the intelligibility of synthetic speech under noisy or low-fidelity acoustic conditions. Techniques explored improve speech in a natural manner, such that training won't be required for the user to understand the enhanced speech. While the improvements are natural in this respect, the changes aren't limited to creating only speech that is achievable by a human vocal tract. Modifications fall into three broad classes: increasing phoneme amplitude, altering spectral shape, and lengthening phoneme duration. Listening tests conducted in noisy and noise-free conditions demonstrate significant improvements to intelligibility for most of the subject phonemes.

1. MOTIVATION

We foresee a rapidly increasing need for high-intelligibility speech synthesizers. As computer systems become more miniaturized and mobile, there will be more circumstances where video displays are impractical. Whether because the computing device is too small for an adequate visual display, such as in a cell phone, or because it would be dangerous to divert an operator's visual attention, as when driving an automobile, speech synthesis is an attractive alternative user interface. Unfortunately, many portable systems are typically used under acoustically challenging conditions, for example within the noisy environment of an automobile, or subject to the limited acoustic output and low fidelity of tiny, low-powered speakers. In these cases, it is much more important that the speech be understood than it is for the speech to sound natural. Our goal is to spur the development of synthesizers that maximize intelligibility, which is analogous to how modern fonts maximize the legibility of text.

2. BACKGROUND

Arguably work in the area of speech intelligibility spans nearly two centuries. Early work focused on hearing aids [1]. Subsequent studies focused on the intelligibility of specific phonemes [2,3], while recent work studied intelligibility in a more general sense [4,5]. With the advent of speech synthesizers, researchers have examined the intelligibility of synthesized speech under both clear and noisy conditions [7,8,9].

Time-honored work has been done on methods of improving speech intelligibility. Since World War II, people have known that speech intelligibility could be improved by increasing the amplitude of consonants relative to that of vowels [6]. Humans also tend to increase the consonant-vowel energy ratio when

asked to speak clearly [10]. In addition, studies have shown that humans instinctively change the way they speak to improve intelligibility. When speaking in a noisy environment, humans tend to automatically alter their speech. This phenomenon is known as the Lombard Effect. Lombard speech has been found to increase speech intelligibility in noise in many cases [11].

Information from these studies allows us to improve intelligibility in two ways. We both use traditional signal-processing approaches as well as mimic what humans do to improve intelligibility.

3. OVERVIEW

For this paper we explored methods of improving the intelligibility of synthesized speech in noise. We focused on methods that do not require user training to achieve the improved intelligibility. We based our experiments on a DECTalk formant synthesizer [12], version 4.6. This synthesizer allows us to identify phonemes and extract associated acoustic parameters. This facilitates modification of the synthetic speech output. DECTalk has the additional advantages of a high initial intelligibility [7,8] and a low-complexity, memory-efficient realization. It is thus well suited to portable applications. While our study focused on a specific synthesizer and language, we kept our approach general so that it can be applied to any synthesizer or language. We first determined the phonemes that need the most improvement in intelligibility, a set of consonants. Next we developed tests to measure the intelligibility of these consonants in the presence of noise. We then proceeded in three phases:

1. Determine whether amplification of the selected consonants is sufficient to improve their intelligibility in noise.
2. For the consonants that do not respond to amplification, determine if a change in spectral shape improves intelligibility.
3. For the consonants that do not respond to the above two techniques, determine if time-stretching the consonant improves intelligibility.

4. TARGET CONSONANTS

In order to limit the scope of our work, we targeted the least intelligible phonemes in the American English language. It is widely known that consonants are less intelligible than vowels [5]. This is not surprising since consonant sounds are generally both weaker in strength and shorter in duration than vowels. Figure 1 shows a plot of the average RMS value versus the duration in milliseconds for the 55 DECTalk phonemes. These

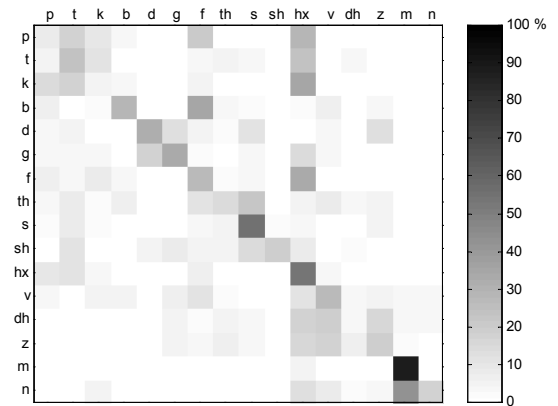
5.2. Results

The first phase of our study was an extension of previous work that improved intelligibility by indirectly increasing the amplitude of the consonants relative to the vowels, either by clipping or by fast limiting the speech signal [6]. We amplified each consonant by 600% or until the signal level was just below clipping, whichever was less. Half of our set of consonants improved with just amplification: *s*, *sh*, *m*, *n*, *t*, *k*, *b*, and *d*. For the remaining eight consonants this method provided little improvement and in some cases actually introduced more errors. For instance, after amplification, *hx*'s were more easily mistaken for *p*'s, and *v*'s were more easily mistaken for *m*'s.

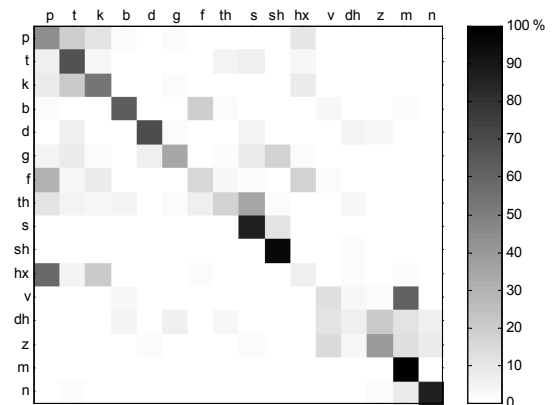
The second phase of our study was designed to test whether the lack of intelligibility improvement for the remaining consonants was due to poor phonetic reproduction by the DECTalk synthesizer. We replaced each of these consonants with an amplified recording of that consonant produced by a male speaker trying to speak the associated syllable clearly. The amount of amplification was comparable to that used in the first phase. Five consonants, *p*, *g*, *hx*, *z*, and *v*, had a significant improvement in intelligibility using this approach. That there was little improvement to the *f*, *th*, and *dh* phonemes is not too surprising. For natural speech, these fricatives are weaker than the other fricatives [13]. The output of the DECTalk synthesizer is consistent with this observation. Figure 1 shows that *f* and *th* are the two weakest consonants we tested. Unamplified, these consonants are inaudible for the noise level we used. Unfortunately with amplification these consonants no longer sound as expected. In particular, the amplified *dh* sounded more like a *v* or a *z*. Our difficulties in improving the intelligibility of the *th* and *dh* consonants may also be partially due to the limited bandwidth of the DECTalk synthesizer. According to notes on spectrogram reading, the major energy onset in the frequency domain for these consonants starts at 6 kHz for an adult male [13]. Perhaps increasing the sampling rate of the synthesizer from 11.025 kHz to 16 kHz would help.

In the third phase we attempted to improve the intelligibility of consonants that apparently couldn't be improved by amplification or the substitution of human clear speech utterances. Studies of human clear speech indicate that people tend to increase the voicing onset time and increase the duration of frication noise [10]. We therefore tried doubling the duration of the consonants, *f*, *th*, and *dh*. We enhanced the other consonants as in phase 2. There was no significant improvement in intelligibility using this technique for *dh* and *th*. This was expected because the time-stretching aspect was already captured in phase 2 by using clips of human clear speech. The intelligibility of *f* unexpectedly improved from an error rate of 62% to 44%.

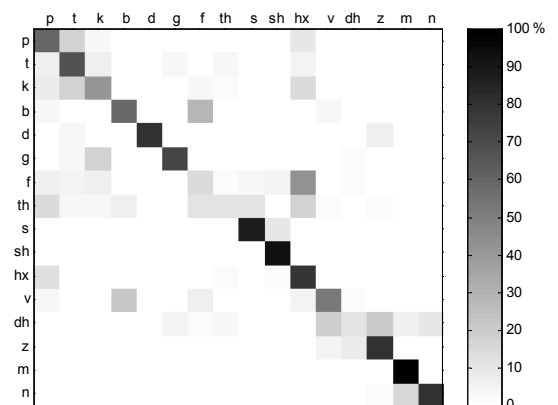
Figure 2 summarizes the overall outcome of our tests in the form of confusion matrices. The actual consonants are labeled on the left while the perceived consonants are labeled across the top. The count in each entry, as a percentage of the total, is depicted according to the shading scale on the right. Detailed scores in the form of Excel spreadsheets are available at the following [web site](#):



(a)



(b)



(c)

Figure 2: Confusion matrices for (a) original data set, (b) enhanced data set using only amplification, and (c) enhanced data set using both amplification and human clear speech replacement for certain consonants

<http://crl.research.compaq.com/projects/speechsynth>. Figure 2(a) shows the intelligibility results of the unmodified data set. The unmodified set scored an overall error rate of roughly 71%. It is satisfying to note that, of the consonants we tested, those with the highest intelligibility, *m* and *s*, are located highest and widest on the phoneme-power-versus-duration curve in Figure 1, while those with the lowest intelligibility, *p* and *k*, were located closest to the origin. Figure 2(b) shows the results for the amplified DECTalk consonants. Amplification alone improved the overall error rate to 49%. Figure 2(c) shows the results for the enhanced data set. This set combines amplification for most of the consonants with human consonant replacement and amplification for the phonemes: *p*, *g*, *hx*, *z*, *v*, *f*, and *th*. The enhanced consonants scored an overall error rate of less than 37%. It is clear that the enhancements have improved the intelligibility of the synthesized speech.

Finally, to check if the enhancements to improve speech intelligibility in noise affected intelligibility without noise we conducted a test of the phase 2 speech enhancements without the interfering noise. Test results as shown in Figure 3 indicate that, in the noise-free case, except for *g* and *hx*, intelligibility for the enhanced consonants was higher than or roughly equal to the intelligibility of the original DECTalk consonants. Both these enhanced consonants were more often mistaken for a *k* in the noise-free case.

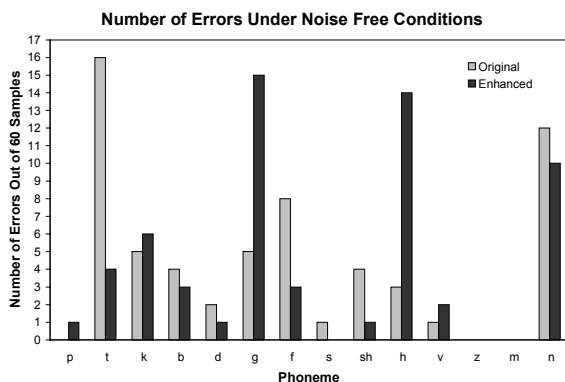


Figure 3: Noise-Free Error Statistics

6. FUTURE WORK

Our preliminary work in this area shows much promise. There is still much work to be done. We should isolate and identify the clear speech features that contributed to the increase in intelligibility of *p*, *g*, *hx*, *z*, and *v*. We should also conduct intelligibility tests with noise at other levels and other spectral distributions. This study only examined word-initial consonants; intelligibility enhancements to word-medial, word-final consonants should also be studied. More traits of human clear speech [4,10] should be incorporated in the synthesizer to see if they improve intelligibility. These include: extending formant transitions, including sound insertions (a schwa vowel) after voiced consonants, slowing down the speech rate, expanding

the first and second formant ranges to make the vowels *aa*, *iy*, *ow* more distinct (enlarge the *aa*, *iy*, *ow* vowel triangle).

7. REFERENCES

- Berger, K.W., (1974), "The hearing aid: its operation and development," National Hearing Aid Society, Livonia, MI.
- Miller, G.A. & Nicely, P.E, (1955), "An Analysis of Perceptual Confusions Among Some English Consonants," J. Acoust. Soc. Am., vol. 27, no. 2, 338-352.
- Fairbanks, G., (1958), "The Test of Phonemic Differentiation: The Rhyme Test," J. Acoust. Soc. Am., vol. 30, no. 7, 596-600.
- Bradlow, A.R., Torretta, G.M., & Pisoni, D.B., (1996), "Intelligibility of Normal Speech I: Global and Fine-grained acoustic-phonetic talker characteristics," Speech Communication, vol. 20, no. 3-4, 255-272
- Neel, A.T., Bradlow, A.R., & Pisoni, D.B., (1996), "Intelligibility of Normal Speech II: Analysis of Transcription Errors," Research on Spoken Language Processing, Progress Report No. 21.
- Kretsinger, E.A. & Young, N.B., (1960), "The Use of Fast Limiting to Improve the Intelligibility of Speech in Noise," Speech Monogr. No. 27, 63-69.
- Wright, J.T., B.J. Malsheen & M. Peet, (1986), "Comparison of segmental intelligibility and pronunciation accuracy for two commercial text-to-speech systems," Proc. AVIOS, 235-261.
- Logan J., Greene B., & Pisoni D. (1989). "Segmental Intelligibility of Synthetic Speech Produced by Rule", J. Acoust. Soc. Am., vol. 86, no. 2, 566-581.
- Simpson, C.A. & Navarro, T. (1984), "Intelligibility of Computer Generated Speech as a Function of Multiple Factors," Proc. IEEE 1984 Nat. Aerospace and Elec. Conf., vol. 2, 932-940.
- Pichney, M.A., Durlach, N.I., & Braida, L.D., (1986), "Speaking Clearly for the Hard of Hearing II: Acoustic Characteristics of Clear and Conversational Speech," J. Speech and Hearing Research, 29, 434-446.
- Junqua, J-C, (1993), "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers," J. Acoust. Soc. Am., vol. 93, no. 1, 510-524.
- Klatt, D.H. & Klatt, L.C., (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., vol. 87, no. 2, 820-857.
- Zue, V. (1985), "Notes on Spectrogram Reading", Mass. Inst. Tech. Course 6.67s lecture notes, Cambridge, MA.