Loudmouth: Modifying Text-to-Speech Synthesis in Noise

Rupal Patel
Dept. of Speech Language Pathology
Northeastern University
Boston, MA 02115
+ 1 617 373 5842
r.patel@neu.edu

Michael Everett
Dept. of Computer & Information Science
Northeastern University
Boston, MA 02115
everettm@ccs.neu.edu

Eldar Sadikov
Dept. of Computer & Information Science
Northeastern University
Boston, MA 02115
eldar@ccs.neu.edu

ABSTRACT

Current speech synthesis technology is difficult to understand in everyday noise situations. Although there is a significant body of work on how humans modify their speech in noise, the results have yet to be implemented in a synthesizer. Algorithms capable of processing and incorporating these modifications may lead to improved speech intelligibility of assistive communication aids and more generally of spoken dialogue systems. We describe our efforts in building the Loudmouth synthesizer which emulates human modifications to speech in noise. A perceptual experiment indicated that Loudmouth achieved a statistically significant gain in intelligibility compared to a standard synthesizer in noise.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences - *Health*

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Speech Synthesis, Text-to-Speech Synthesis (TTS), Augmentative and Alternative Communication (AAC)

1. INTRODUCTION

Text-to-speech (TTS) synthesis is increasingly being integrated into everyday appliances and services such as directory assistance, public transportation announcements and for automated banking. In addition, this technology serves an essential need for individuals with severe speech and motor impairments. Many of these individuals rely on speech synthesis installed on their assistive aid as a primary means of communication.

Despite the proliferation of applications using TTS synthesis, few efforts have been aimed at improving speech intelligibility in everyday noise situations. As noise levels rise, increasing the volume alone is not sufficient for achieving acceptable levels of intelligibility. More often, increasing the volume distorts the signal and actually degrades the overall intelligibility.

The Lombard effect is a change to speech produced in noise [1, 2, 3]. Studies have shown that in noisy environments people tend to increase their intensity, fundamental frequency (f0) and duration [2, 3, 4], collectively referred to as prosody, in order to improve speech intelligibility [5,6].

Langner and Black [7, 8] attempted to harness the Lombard Effect by building a concatenative synthesizer using diphones recorded in noise. They reported improved TTS intelligibility for a limited domain task. However, their modifications did not consider the semantic content of the spoken utterances.

Given that changes in intensity, f0, and duration are also used to convey linguistic stress and intention [9, 10], it is likely that modifications to speech in noise may be differentially applied based on linguistic content. Recent work suggests that the Lombard effect is enhanced for semantically salient words within a sentence compared to non-salient words [11, 12]. These empirical findings led to the Loudmouth synthesizer which incorporates acoustic and linguistic alterations to speech in noise.

2. SYSTEM OVERVIEW

Loudmouth was implemented by modifying FreeTTS, an open source concatenative TTS synthesizer [13]. Standard concatenative synthesizers are comprised of front and back-end processors. The front-end prepares the inputted text for synthesis by discovering all the diphone pairs for the sentence, while the back-end calculates the duration and the frequency contour for the diphones prior to synthesis. The FreeTTS front and back-end were modified in Loudmouth. A Text Analyzer was added to the front-end to allow for identification of linguistic content. In the back-end, the Durator was modified to manipulate word duration, the Contour Generator was modified to alter f0, and the Concatenator was modified to amplify intensity.

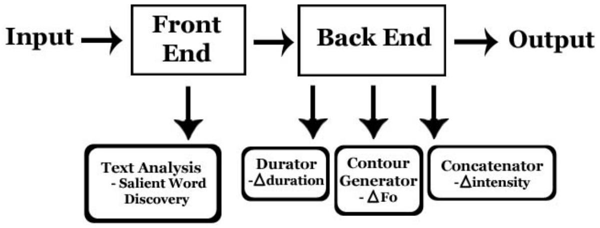

Figure 1. Overview of the Loudmouth TTS modifications.

ASSETS '06, October 22–25, 2006, Portland, Oregon, USA.
ACM 1-59593-290-9/06/0010.

2.1 Text Analysis

An input sentence is operated on by the Text Analyzer to locate semantically salient words within the utterance. A part-of-speech tagger is used to identify salient parts of speech such as nouns, verbs, and other content words as well as non-salient parts of speech such as the function words "of", "a", "the". Each part of speech is associated with a default increase in duration, f0, and intensity based on empirical data from modifications male speakers made in noise [2, 5, 10, 11, 12]. Thus the Text Analyzer outputs each word with an associated duration, f0, and intensity value based on its part of speech.

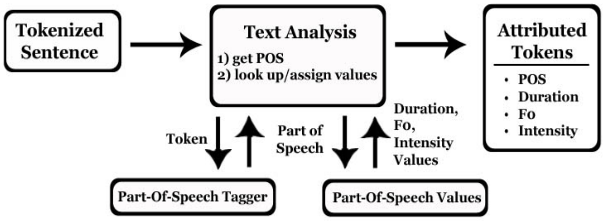

Figure 2. Text Analysis of an input sentence.

2.2 Acoustic Modifications

2.2.1 Duration

In standard synthesis, the duration of each diphone is calculated to determine its length. In Loudmouth these values are modified based on the duration assigned during Text Analysis. Salient words are assigned larger values than non-salient words to lengthen the duration of each diphone. Salient words averaged 546ms in duration, an increase of 98ms over non-salient words.

2.2.2 Fundamental Frequency

To alter the pitch of each word the f0 contour of the standard synthesizer is shifted by an integer value based on the output of Text Analysis. The average f0 of salient words was approximately 20 Hertz greater than non-salient words.

2.2.3 Intensity

The perceived loudness of the synthesizer is modified by altering the intensity of each word based on the output of Text Analysis. Intensity modifications are performed as the audio byte stream is created for each word. The byte stream of salient words is multiplied by a larger value than that of non-salient words. Salient words averaged 4dB higher in intensity than non-salient words.

3. PERCEPTUAL EXPERIMENT

A perceptual experiment examined the intelligibility between Loudmouth and the unmodified FreeTTS synthesizer in the presence of noise using the defualt male voice of FreeTTS. The study consisted of ten adult monolingual speakers of English (5 male, 5 female; mean age = 21.3 years old). The participants engaged in an interactive computer game in which they had to move characters on the screen based on commands issued by the two synthesized voices.

The study was run in two conditions, silence and 80dB of multi-talker noise. In silence, Loudmouth and the unmodified synthesizer showed almost 100% correct word recognition suggesting that our modifications did not degrade overall voice quality. In noise, Loudmouth was 7% more intelligible than the unmodified synthesizer. Thus the results suggest that the modifications implemented in Loudmouth may be a viable approach for improving TTS intelligibility in noise.

4. DEMONSTRATION

This demonstration will introduce Loudmouth and allow participants to interactively synthesize speech. Participants will be able to hear and visually observe the output of Loudmouth compared to that of a standard synthesizer. A graphical user interface will also allow participants to further customize the Loudmouth output by selectively manipulating f0, intensity and duration of each diphone.

5. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under grant No. 0509935.

6. REFERENCES

[1] Lombard, E., "Le signe l'élévation de la voix", *Maladies Oreille, Larynx, Nez, Pharynx, 27*, 101-119, 1911.

[2] Junqua, J., "The Lombard reflex and its role on human listeners and automatic speech recognizers", *J. of the Acoustical Society of America 93 (1)*, 510-524, 1993.

[3] Junqua, J., "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication, 20 (1-2)*, 13-22, 1996.

[4] Brown, W., Brandt, J., and John, F., "The effect of masking on vocal intensity during vocal and whispered speech", *J. of Auditory Research, 12 (2)*, 157-161, 1972.

[5] Letowski, T., Frank, T., and Caravella, J., "Acoustical properties of speech produced in noise through supra-aural headphones", *Ear and Hearing*, 14 (5), 332-338, 1993.

[6] Lane, H., and Tranel, B., "The Lombard sign and the role of hearing in speech", *J. of Speech and Hearing Research*, 14, 677-709, 1971.

[7] Langner, B. and Black, A. "Creating a Database of Speech in Noise for Unit Selection Synthesis", *ISCA Workshop on Speech Synthesis*, Pittsburgh, PA 2004.

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

[9] Cutler, A., "Segmentation problems, rhythmic Solutions", *Lingua*, 92, 81-104, 1994.

[10] Rivers, C. and Rastatter, P., "The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults", *The Journal of Auditory Research*, 25 (1), 37-45, 1985.

[11] Patel, R., and Syeda, M., "The Influence of Semantic Information on the Acoustics of Speech in Noise", *Conference of the Acoustical Society of America*, NY, NY, 2004.

[12] Patel, R., Whang, J., and Nunez, P.,"Prosodic Alterations to Speech Produced in Noise", *Speech Motor Control / Motor Speech Disorders Conference*, Albuquerque, NM, 2004.

[13] Walker, W., Lamere, P., and Kwok, P. "Introduction", http://freetts.sourceforge.net, 2001.