An Overview of the VUB Entry for the 2013 Hurricane Challenge

Henk Brouckxon\textsuperscript{1,2}, Werner Verhelst\textsuperscript{1,2}

\textsuperscript{1} Vrije Universiteit Brussel, Dept. ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium
\textsuperscript{2} iMinds, Dept. of Future Media and Imaging, G. Crommenlaan 8, B-9050 Ghent, Belgium

\texttt{hbrouckx@etro.vub.ac.be, wverhels@etro.vub.ac.be}

Abstract

This paper describes the \textit{SINCoFETS} entry for the Hurricane challenge\cite{1}, in which intelligibility enhancement algorithms for speech presentation in noise are compared. The proposed system combines noise-independent non-uniform time scaling and dynamics compression algorithms with noise-dependent frequency equalization to improve the robustness of speech intelligibility against noise. The algorithms in the system are described and a short discussion of the results is given.

\textbf{Index Terms}: speech intelligibility, non-uniform time scaling, frequency equalization

1. Introduction

When speech is presented through a public addressing (PA) system, the background noise in the presentation environment can have a large impact on its intelligibility. By applying pre-processing algorithms like high-pass filtering\cite{2} and/or dynamics compression\cite{3}, the robustness of speech intelligibility against noise can however be improved. This paper describes our entry \textit{(SINCoFETS)} for the Hurricane challenge\cite{1}, in which intelligibility enhancement algorithms were subjectively evaluated and compared for different noise types and signal-to-noise (SNR) ratios.

2. Algorithm

In a previous paper\cite{4} we proposed an intelligibility enhancement algorithm based on time- and frequency-dependent equalization of speech. The \textit{SINCoFETS} system combines this algorithm with two noise-independent modification algorithms that work on complementary properties of the speech signal. The complete system is shown in figure 1 and described in the following sections. Where possible, the algorithm settings are chosen with an emphasis on retaining high quality and a high degree of naturalness in the processed speech.

2.1. Non-uniform time scaling

In a noisy environment, human speakers adapt their way of talking to improve intelligibility, a phenomenon known as the Lombard effect\cite{5}. One typical adaptation is a decreased speaking rate, giving the listener more time to understand the speech. This decrease is non-uniform, and typically shows slower speaking rates at speech sounds that are important or difficult to understand. In \textit{SINCoFETS}, we apply a similar time-scaling strategy where consonants are slowed down more because they are typically most susceptible to noise interference. To this end, the non-uniform time-scaling algorithm in figure 2 was implemented. The \textit{Consonant/Vowel/Pause detector} classifies all speech sounds\cite{6}. Vowels are e.g. detected using maxima in the mel-scaled Reduced Energy Cumulative Function\cite{7} and pauses are detected based on the long term spectral estimation (LTSE) and long term spectral divergence (LTD). The \textit{Time-Scaling Factors} block determines a suitable time scaling factor for each speech sound, based on the classification and predefined time-scaling factors for each type of sound. The \textit{Non-Uniform Time-Scaling} block applies the time scaling factors to the speech signal, using high-quality WSOLA (Waveform-Similarity based OverLap Add)\cite{8,9}.

Overall, sentences were slowed down as much as possible within the constraints of the Hurricane challenge. Consonants were additionally slowed down by a factor 0.6, and pauses were sped up by an additional factor 1.2.

2.2. Dynamics compression

Public addressing systems use slow-acting dynamics compression to compensate for sentence-level amplitude variations caused by breathing or inter-speaker differences\cite{10,11}. On a shorter timescale however, large amplitude differences also exist between (strong) vowels and (weaker) consonants. Due to these differences, environmental noise can be detrimental to consonant audibility even at SNRs for which vowels remain clearly audible. By implementing a fast-acting level detector that detects level changes between vowels and consonants, the dynamics compressor can redistribute the speech energy between these speech sounds more evenly.

As shown in figure 3 the compressor’s sidechain measures the input signal level (Level Detector), and determines a gain (Gain Characteristic) that stabilizes the speech signal level (G). The gain characteristic defines the correspondence in dB scale between the levels of the input and output signals as in figure 4. A delay block is included in the forward chain to compensate

Figure 1: The \textit{SINCoFETS} system (N-blocks indicate renormalisation to the original RMS signal level)

Figure 2: The Non-uniform Time-Scaling algorithm
Figure 3: The dynamics compressor

Figure 4: Gain Characteristic of the dynamics compressor

2.3. Noise-dependent frequency equalization

Psycho-acoustic research [12, 13] has shown that a threshold Sound Pressure Level (SPL) exists for each frequency, below which the human hearing system does not perceive any sound. In the presence of background noise this hearing threshold is increased, causing the noise to ‘mask’ the presented speech. Psycho-acoustical models, like the ones used in MP3 [13] and AAC encoders, predict this effect and provide a Signal-to-Masking Ratio (SMR) for the frequency components in the speech. The frequency regions close to the first three speech formants, which the human hearing system does not perceive any sound, are known to be most important for intelligibility [14]. In speech, the frequency regions close to the first three formants, based on LPC pole tracking [16].

- A formant tracker for the first three formants, based on LPC pole tracking [16]
- Based on the measured and desired SMR for the formants, The Gain Calculation block determines a suitable gain and tuning frequency for the parametric equalizers.
- The ‘Gain Smoothing’ block limits fast changes in the gain factors and tuning frequencies to avoid artifacts.

Table 1: Results of the subjective test, $CS = \text{Competing Speaker}, SSN = \text{Speech Shaped Noise}$

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR</th>
<th>Original</th>
<th>Processed</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>-7 dB</td>
<td>85.1 +/- 1.5</td>
<td>88.1 +/- 1.1</td>
<td>3.0</td>
</tr>
<tr>
<td>CS</td>
<td>-14 dB</td>
<td>57.0 +/- 2.4</td>
<td>59.9 +/- 2.3</td>
<td>2.9</td>
</tr>
<tr>
<td>CS</td>
<td>-21 dB</td>
<td>24.8 +/- 1.9</td>
<td>23.2 +/- 1.8</td>
<td>-1.7</td>
</tr>
<tr>
<td>SSN</td>
<td>+1 dB</td>
<td>88.3 +/- 1.3</td>
<td>87.6 +/- 1.2</td>
<td>0.7</td>
</tr>
<tr>
<td>SSN</td>
<td>-4 dB</td>
<td>63.0 +/- 2.2</td>
<td>74.5 +/- 1.8</td>
<td>11.5</td>
</tr>
<tr>
<td>SSN</td>
<td>-9 dB</td>
<td>17.3 +/- 1.8</td>
<td>23.6 +/- 2.0</td>
<td>6.3</td>
</tr>
</tbody>
</table>

The research described in this paper was performed with the support of the iMinds TRACK and RAILS projects. The iMinds TRACK and RAILS projects are cofunded by iMinds (Interdisciplinary institute for Technology), a research institute founded by the Flemish Government, and with project support of IWT. We also thank Mike Demol for his help with the time-scaling software.

3. Implementation and results

The $SINCoFETS$ system was implemented in Matlab and applied to 180 sentences for six different noise backgrounds (Speech Shaped Noise (SSN) at -9, -4 and +1 dB SNR, and Competing Speaker (CS) at -21, -14 and -7 dB SNR). Table 1 summarises the results for the $SINCoFETS$ system [1]. It shows the percentage of keywords that were correctly understood (+/- variance) for the original and processed speech, and their differences (intelligibility Gain).

A more detailed description of this system is given in [4].
5. References


