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} I-Minds, Dept. of Future Media and Imaging, G. Crommenlaan 8, B-9050 Ghent, Belgium

hbrouckx@etro.vub.ac.be, wverhels@etro.vub.ac.be

Abstract

This paper describes the SINCoFETS entry for the Hurricane challenge [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.

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 preprocessing algorithms like high-pass filtering [2] and/or dynamics compression [3], the robustness of speech intelligibility against noise can however be improved. This paper describes our entry (SINCoFETS) for the Hurricane challenge [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 [4] we proposed an intelligibility enhancement algorithm based on time- and frequency-dependent equalization of speech. The 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 [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 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 Consonant/Vowel/Pause detector classifies all speech sounds [6]. Vowels are e.g. detected using maxima in the mel-scaled Reduced Energy Cumulative Function [7] and pauses are detected based on the long term spectral estimation (LTSE) and long term spectral divergence (LTSD). The Time-Scaling Factors block determines a suitable timescaling factor for each speech sound, based on the classification and predefined time-scaling factors for each type of sound. The Non-Uniform Time-Scaling block applies the time scaling factors to the speech signal, using high-quality WSOLA (Waveform-Similarity based Overlap Add) [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 [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 SINCoFETS system (N-blocks indicate renormalisation to the original RMS signal level)

Figure 2: The Non-uniform Time-Scaling algorithm
for the level detector’s latency.

The SINCofETS dynamics compressor was implemented based on a fast-acting first-order level detector:

\[
level(n) = a, |x(n)| + (1 - a).level(n - 1)
\]

\[
a = a_{AT} = e^{-Ts/T_{AT}} \text{ if level}(n - 1) < |x(n)|
\]

\[
a = a_{RE} = e^{-Ts/T_{RE}} \text{ if level}(n - 1) > |x(n)|
\]

with \(a_{AT}\) and \(a_{RE}\) the attack and release time constants of the level detector and \(T_s\) the sampling period of the digital system. For the SINCofETS system, a fast acting attack time \(T_{AT} = 1.0\text{ms}\) and release time \(T_{RE} = 15.0\text{ms}\) were chosen.

### 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 are known to be most important for intelligibility [14]. In [4] we proposed a system (figure 5) in which parametric equalizers [15] are used to raise these formants above the hearing threshold using minimal overall amplification:

- **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.

### 4. Acknowledgements

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.
5. References


