An Overview of the VUB Entry for the 2013 Hurricane Challenge

Henk Brouckxon[1,2], *Werner Verhelst*[1,2]

[1] Vrije Universiteit Brussel, Dept. ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium
[2] iMinds, 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 pre-processing 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 time scaling 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

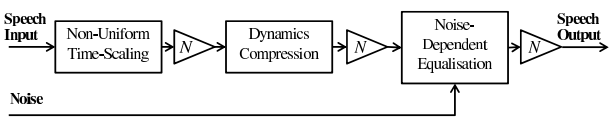

Figure 1: The $SINCoFETS$ system (N-blocks indicate re-normalisation to the original RMS signal level)

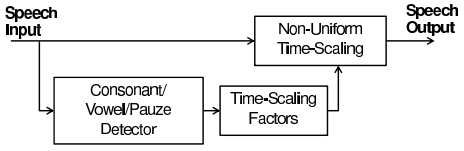

Figure 2: The Non-uniform Time-Scaling algorithm

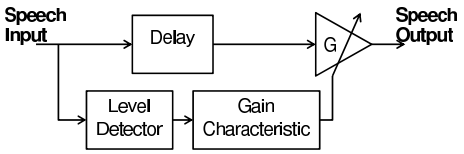

Figure 3: The dynamics compressor

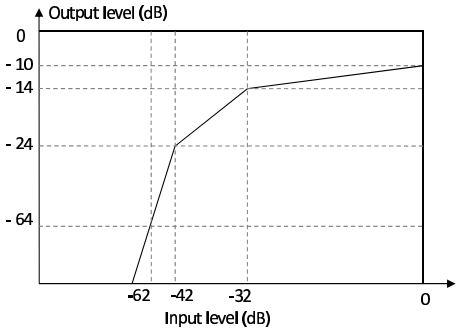

Figure 4: Gain Characteristic of the dynamics compressor

for the *level detector*'s latency.

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

$$level(n) = a.|x(n)| + (1-a).level(n-1) \quad (1)$$

$$a = a_{AT} = e^{-Ts/t_{AT}} \;\; if \;\; level(n-1) < |x(n)| \quad (2)$$

$$a = a_{RE} = e^{-Ts/t_{RE}} \;\; if \;\; level(n-1) > |x(n)| \quad (3)$$

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 $t_{AT} = 1.0ms$ and release time $t_{RE} = 15.0ms$ 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]
- A psycho-acoustical model (*Speech analysis*, *Noise Analysis* and *Audibility Evaluation* blocks) based on [13] determines the audibility of the formants.
- 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.

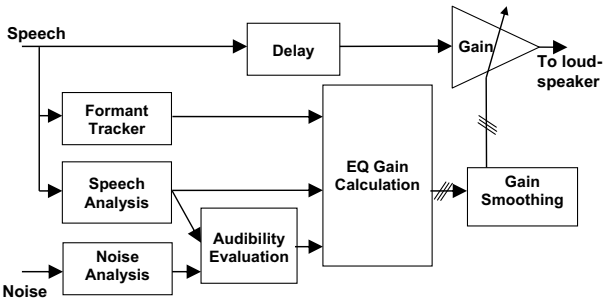

Figure 5: The noise-dependent frequency equalization system

- Three parametric equalizers (*Gain*) apply the desired time-varying equalization to the speech signal.

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

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).

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

Table 1: Results of the subjective test, CS = Competing Speaker; SSN = Speech Shaped Noise

For Competing Speaker (CS) noise, the scores show relatively small (and statistically less significant) differences. This result can be explained to some degree by the highly dynamic nature of the noise, for which the system's parameters were not optimised. It is expected that, if more agressive parameter settings would be used, better intelligibility could be obtained at the cost of some speech distortion. For Speech Shaped Noise (SSN) a clear (and statistically significant) improvement in intelligibility is seen for the two lowest SNRs. The more stationary nature of this background is more similar to the PA application for which the system parameters were chosen.

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

[1] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the hurricane challenge," in *Proceedings of Interspeech 2013*, 2013.

[2] J.D.Griffiths, "Optimum linear filter for speech transmission," *Journal of the Acoustical Society of America*, vol. 43, no. 1, pp. 81–86, 1968.

[3] K.S.Rhebergen, N.J.Versfeld, and W.A.Dreschler, "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise," *Journal of the Acoustical Society of America*, vol. 126, no. 6, pp. 3236–3245, 2009.

[4] H. Brouckxon, W. Verhelst, and B. De Schuymer, "Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments," in *proceedings of Interspeech 2008*, Brisbane, Australia, 2008.

[5] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[6] M. Demol, W. Verhelst, and P. Verhoeve, "A study of speech pauses for multilingual time-scaling applications," in *proceedings of ISCA-ITRW Multiling 2006*, Stellenbosch, South Africa, 2006.

[7] F. Pellegrino and R. Andre-Obrecht, "Automatic language identification: an alternative approach to phonetic modelling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[8] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, vol. 2. IEEE, 1993, pp. 554–557.

[9] M. Demol, W. Verhelst, K. Struyve, and P. Verhoeve, "Efficient non-uniform time-scaling of speech with wsola," in *proceedings of Speech and Computers 2005 (SPECOM-2005)*, Patras, Greece, 2005.

[10] H. Brouckxon, W. Verhelst, K. Struyve, and P. Verhoeve, "Design and evaluation of a microphone signal conditioning system," in *proceedings of ELMAR-2005*, Zadar, Croatia, 2005.

[11] U. Zolzer, *DAFX - Digital Audio Effects, Chapter 5*. John Wiley & Sons, 2002.

[12] H. Fletcher and W. Munson, "Loudness, it's definition, measurement and calculation," *Journal of the Acoustical Society of America*, vol. 5, pp. 82–108, 1933.

[13] "ISO/IEC standard 226:2003, Acoustics - Normal equal-loudness-level contours," 2003.

[14] G. Fant, *Acoustic theory of speech production*. Mouton, 's Gravenhage, Nederland, 1960.

[15] P. Regalia and S. Mitra, "Tunable digital frequency response equalisation filters," *IEEE Transactions On Acoustics, Speech and Signal Processing*, vol. ASSP-35, no. 1, pp. 118–120, 1987.

[16] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE transactions on Acoustics, Speech and Signal Processing*, April 1974.