Statistical Synthesizer with Embedded Prosodic and Spectral Modifications to Generate Highly Intelligible Speech in Noise

D. Erro [1,2], *T.C. Zorilă* [3], *Y. Stylianou* [4], *E. Navas* [1], *I. Hernaez* [1]

[1] AHOLAB, University of the Basque Country, Bilbao, Spain
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[3] POLITEHNICA University of Bucharest, Bucharest, Romania
[4] FORTH & University of Crete, Heraklion, Greece

derro@aholab.ehu.es

10.21437/Interspeech.2013-765

Abstract

This paper describes a statistical parametric speech synthesizer that, despite having been trained on an ordinary synthesis database and without any adaptation data, is able to generate highly intelligible speech in noisy environments. By using a simple and flexible vocoder based on a harmonic model, it applies several noise-independent modifications to durations, pitch level and range, energy contour, formant sharpness, and intensity of particular spectral bands. The system has been evaluated by means of a large subjective test, the results of which show that the suggested approach clearly outperforms the reference TTS systems and even unmodified natural speech in some conditions.

Index Terms: statistical parametric speech synthesis, speech intelligibility in noise, speech modification and transformation

1. Introduction

Speech synthesis databases are usually recorded at very high signal-to-noise ratio (SNR) in silent or even anechoic rooms. While being recorded, human speakers unconsciously adapt their voice to this particular environment. Synthetic voices built from the recorded database inherit the acoustic characteristics of the human voice. Then, when the synthesizer is used in noisy conditions (at the airport, for instance) it is often hard for listeners to understand the message. The problem is evident: while human talkers are capable of adapting their voices to the environment, machines are not. It is therefore necessary to develop methods for enhancing the intelligibility of clean synthetic speech before it is played in a noisy environment.

Previously, this problem has been tackled in two different ways: (i) recording a new database in the desired conditions and then using voice conversion, speaker adaptation, inter/extrapolation, or any other statistical mapping technique to transform one style into the other [1][2][3][4]; (ii) using the original clean speech database and applying expert knowledge and signal processing techniques to enhance the output of the synthesizer [5][3]. The main advantage of the latter strategy is indeed the fact that it is costless. It requires, nonetheless, an adequate flexible signal processing framework.

With this goal in mind, in a previous work [6] we showed the usefulness of a simple harmonic model to enhance the intelligibility of clean natural speech in noise. In this paper we go one step beyond: we have developed a hidden Markov model (HMM) based text-to-speech (TTS) synthesizer (see [7] for a review on this widespread synthesis technology) involving a harmonic vocoder which provides both a reliable and effective way of parameterizing and reconstructing speech and a high degree of flexibility for different types of manipulations. Several deterministic noise-independent modifications are performed at spectral and prosodic level: controlled signal lengthening, pitch and pitch range upward modification, dynamic range compression applied to the energy contour, voicing-dependent postfiltering, and enhancement of specific spectral bands. The resulting system has been evaluated together with many others in the context of the Hurricane Challenge [8]. The participants of this international evaluation campaign were provided with a corpus of recorded sentences along with separate noise signals at different SNR conditions. The task to be accomplished was to modify the natural or synthetic speech in such a way to promote its intelligibility, while meeting some constraints on duration and keeping the SNR unaltered. The results achieved by our system confirm the usefulness of the harmonic model and show that the simplicity of the modifications made is not at odds with their effectiveness. In the remainder of this paper, the proposed system is described in detail and the evaluation results are shown and discussed.

2. Synthesizer and Vocoder

The system is based on version 2.1.1 of HTS, the well known open-source HMM-based speech synthesis system [9]. HTS models the acoustic feature vectors provided by a vocoder by means of context-dependent 5-state left-to-right hidden semi Markov models (HSMMs) [10]. In this particular case, a speaker-dependent system was trained from 2863 short utterances at 16 kHz sampling frequency which were provided by the Hurricane Challenge organizers. The context labels used to feed the statistical engine and their corresponding questions for the context-clustering trees were provided by the organizers as well. They had been generated according to the Unisyn pronunciation lexicon [11].

Two acoustic feature streams are used: log-f_0 and a 39th order Mel-cepstral representation of the spectral envelope. No explicit excitation-related stream is given as input to the system because the benefits of doing so (slightly better quality [12], basically) vanish when synthetic speech is played in noise. Pitch related information is modeled by means of multi-space distributions (MSD) [13] due to its discontinuous nature, while spectrum is modeled through continuous HSMMs.

Encouraged by the results reported in [6], we use a purely harmonic vocoder to analyze the training data and reconstruct the synthetic waveforms from parameters. This vocoder is a simplified version of the one presented in [14]. The analysis steps performed by the vocoder are the following: (i) pitch detection and binary voicing decision [15]; (ii) least squares based full-band harmonic analysis assuming $f_0 = 100$ Hz in unvoiced frames [16]; (iii) spectral envelope recovery from

harmonic log-amplitudes via interpolation; (iv) calculation of its corresponding cepstrum and translation into Mel-frequency scale [17]. During synthesis, the amplitude and minimum phase of the harmonics are obtained by sampling the log-spectral envelope given by the Mel-cepstral coefficients at multiples of f_0 (100 Hz if unvoiced). In the absence of an explicit excitation model, we assume a mixed excitation given by an energy-dependent maximum voiced frequency; once it is predicted from the local 0^{th} Mel-cepstral coefficient as suggested in [14], the harmonics at higher frequencies are given random phases.

3. Prosodic and Spectral Modifications

Given the flexibility of both the parametric synthesis framework and the vocoder, many types of modifications can be applied to generate more intelligible and noise-robust speech. As shown in Figure 1, the system described in this paper includes noise-independent modifications at three different stages: on the context labels (point 1 in Figure 1), on the parameters generated by the statistical engine (point 2), and on the internal parameters of the vocoder (point 3). All these modifications are described in the next subsections.

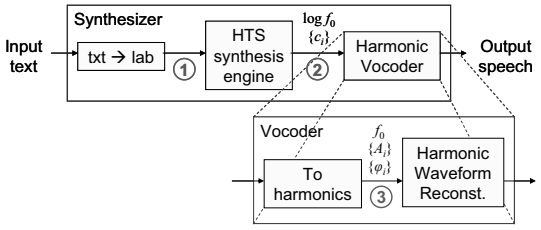

Figure 1: *Block diagram of the proposed system. Modifications are made at points 1, 2 and 3.*

3.1. Duration

Some relationship between speech rate and intelligibility was found in [18][19][20]. In the context of this particular evaluation [8], the duration increments are not arbitrary because the maximal duration is imposed by the noise samples to be combined with the signals. Therefore, we slow down the speech rate as follows:

- After loading the sequence of HSMMs that matches the input context labels, we obtain initial estimates of the phoneme durations by summing the average duration of their five underlying HSMM states.
- All these average phoneme durations are multiplied by the same factor in order to cover the whole available time slot (equal to the duration of the noise samples). In some cases, excesively high factors result in unnatural synthetic speech because the models (particularly the statistics related to the dynamic features) have been trained with speech uttered at faster rates. To alleviate this, we consider an empirically chosen maximum lengthening factor of 1.2.
- We explicitly impose the modified phoneme durations through the input context labels (point 1 in Figure 1) and generate the speech parameters accordingly.
- During generation, we use a slightly modified engine in which the duration of each phoneme is distributed over its HSMM states proportionately to their theoretical average duration, except when the elongation exceeds a given threshold. Beyond this threshold, only the central state is lengthened.

3.2. Pitch level and range

Previous studies reported that artificial f_0 modifications do not produce clear intelligibility improvements [21][22][23]. However, pitch modifications make artificially modified speech more consistent with the way Lombard speech is produced. Human Lombard speech is known to exhibit spectral tilt differences with respect to normal speech [23]. These differences appear when human talkers increase the subglottal pressure, which results in a more pressed phonation. The observed spectral variations are mainly a consequence of such a new glottal excitation. Apart from this, higher subglottal pressure implies more rapid vocal fold vibration, which means that spectral variations are accompanied by higher fundamental frequencies. To make our spectrally modified synthetic speech closer to natural, we artificially introduce this pitch increment effect by means of constant modification factors. According to rough measurements made previously on many voices, we use factor 1.2 for f_0 and factor 1.5 for the standard deviation of log-f_0 throughout the utterance, which is directly related to the f_0 range. To avoid redundant resamplings of the spectral envelope, this is made just after parameter generation and before waveform generation by the vocoder (point 2 in Figure 1).

3.3. Energy contour

Since many low-energy phonemes play a decisive role in intelligibility (plosives, fricatives, vocalic onsets and offsets, nasals [24]), dynamic range compression (DRC) [25] is applied to amplify them at the expense of some energy reduction at high-energy segments. DRC was already shown to play a relevant role in the design of intelligibility enhancement systems dealing with natural speech [26][6]. Nevertheless, the way DRC was applied was slightly different in the two mentioned papers. In [26] it was applied in a sample-by-sample basis, while in [6] constant multiplicative factors were used within each frame. We have chosen the latter strategy because of its higher efficiency and also because it can be controlled at vocoder level (point 3 in Figure 1), before waveform generation.

The procedure can be mathematically described as follows. First, the sum of squared harmonic amplitudes is used as an estimation of the energy at frame k:

$$e^{(k)} = \sum_{i=1}^{I^{(k)}} {A_i^{(k)}}^2 \quad , \quad k = 1 \ldots K \tag{1}$$

where $I^{(k)}$ is the number of harmonics between 0 Hz and the Nyquist frequency, and $\{A_i^{(k)}\}$ are obtained by resampling the Mel-cepstral envelope at multiples of the modified local fundamental frequency [14]. A new set of mapped energies is obtained by means of a non-linear function drc(·) and a correction factor γ that keeps the total energy constant:

$$\hat{e}^{(k)} = \gamma \cdot \mathrm{drc}\left(e^{(k)}\right) \; , \quad \gamma = \frac{\sum_{k=1}^{K} e^{(k)}}{\sum_{k=1}^{K} \mathrm{drc}\left(e^{(k)}\right)} \tag{2}$$

Function drc(·), which is depicted in Figure 2, amplifies low-energy frames (up to 10dB below the maximum) while slightly attenuating high-energy frames. The new energy contour,

$\{\hat{e}^{(k)}\}_{k=1\ldots K}$, will be imposed to the harmonic amplitudes in expression (4), after performing the remaining spectral-level operations.

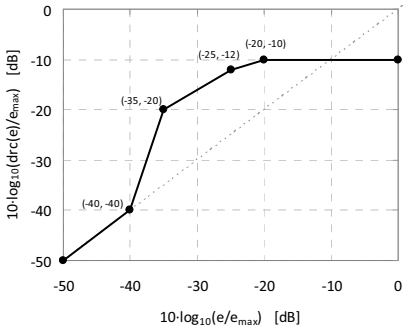

Figure 2: *Energy mapping curve for DRC, where* $e_{\max} = \max\{e^{(1)}, \ldots, e^{(K)}\}$

3.4. Formant sharpness

A postfiltering procedure is applied to sharpen the formant structure. Instead of using a constant postfiltering factor, to avoid the appearance of artifacts, we follow the strategy suggested in [26], where the factor was obtained by multiplying a constant term, 0.25, by the local probability of voicing, $p_v^{(k)}$. Whereas in the mentioned work this required $p_v^{(k)}$ to be estimated from the signal waveform, in HMM-based synthesizers the probability of voicing is one of the internal variables that define the states of the MSD-HSMMs used for log-f_0 contour generation. Therefore, in this system $p_v^{(k)}$ is taken directly from the state sequence during generation, without any extra computational load.

Even though the system generates Mel-cepstral vectors, instead of implementing the postfiltering operation in the cepstral domain as in [26], we operate on the harmonic amplitudes using the implementation described in [27]. This allows a more precise control of the energy (the 0$^{\text{th}}$ Mel-cepstral coefficient is not linearly proportional to the energy) and alleviates inaccuracies in some involved operations such as spectral tilt estimation. Omitting the frame index k for clarity, postfiltering is performed by multiplying each of the I harmonics $\{A_i\}$ by the following term [27]:

$$P_i = \left[A_i \sqrt{\frac{I\left(r_0^2 - 2r_1 r_0 \cos(i\omega_0) + r_1^2\right)}{r_0\left(r_0^2 - r_1^2\right)}} \right]^{0.25\,p_v} \qquad (3)$$

$$r_0 = \sum_{m=1}^{I} A_m^2 \;,\; r_1 = \sum_{m=1}^{I} A_m^2 \cos(m\omega_0)$$

P_i is proportional to the tilt-free harmonic amplitudes. In (3), tilt is estimated and removed implicitly by fitting a 1$^{\text{st}}$-order all-pole filter to the original amplitudes (this requires computing the first two terms of the autocorrelation sequence, r_0 and r_1) and performing inverse filtering [27].

3.5. Intensity

Spectral modifications mimicking intensity increments have been shown to enhance the intelligibility of natural speech in noise [23]. It is logical to assume that this is also true for synthetic speech. Since we are particularly interested in deterministic modifications that can be easily implemented and efficiently applied, we considered two different options during our design: the non-adaptive spectral shaping filter used previously in [26], and the spectral slope modification proposed in [6] (see Figure 3 for a graphical explanation of both). We finally chose the spectral shaping approach for two main reasons: first, it is more consistent with measurements reported in recent works [28][29]; second, it does not produce a significant alteration in the signal quality, while modifying the spectral slope up to the Nyquist frequency may result in local high-frequency artifacts.

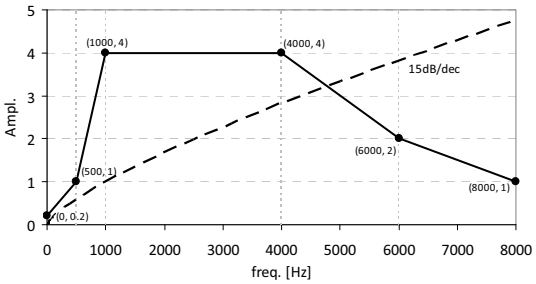

Figure 3: *Spectral shaping filter (solid line) and 15dB/dec spectral slope (dashed line).*

Let us define $H_i^{(k)}$ as the amplitude response of the spectral shaping filter depicted in Figure 3 at the i^{th} harmonic frequency of the k^{th} frame. The final harmonic amplitudes to be used during speech waveform reconstruction, $\{\hat{A}_i^{(k)}\}$, are the result of simultaneously applying postfiltering and spectral shaping on the original ones, $\{A_i^{(k)}\}$, while forcing the energy of the k^{th} modified frame to be equal to $\hat{e}^{(k)}$:

$$\hat{A}_i^{(k)} = \sqrt{\frac{\hat{e}^{(k)}}{\sum_{m=1}^{I^{(k)}} \left(P_m^{(k)} H_m^{(k)} A_m^{(k)}\right)^2}} \cdot P_i^{(k)} H_i^{(k)} A_i^{(k)} \qquad (4)$$

where $\hat{e}^{(k)}$ and $\{P_i^{(k)}\}$ are given by (2) and (3), respectively. No modification is performed on the minimum phases obtained from the Mel-cepstral coefficients [14].

4. Results

As detailed in [8], the described system was evaluated together with several others in the context of the Hurricane Challenge (system identifier: PSSDRC-syn). In the particular case of TTS systems, the training data included 2863 short sentences recorded by a male native British English talker at 16 kHz sampling rate. During the evaluation, TTS systems were given input texts to generate 180 phonetically balanced sentences from the Harvard corpus [30]. After synthesis and/or processing, signals were mixed with two types of noise maskers: speech-shaped noise (SSN) and competing speaker (CS) noise, at three different SNR values each: –9dB, –4dB and 1dB for SSN; –21 dB, –14 dB and –7dB for CS. A perceptual test was carried out at CSTR, Univ. of Edinburgh, in which 175 participants with British English as native language listened to synthetic and reference utterances mixed with noise in sound-isolated booths and were asked to type what they heard. They were not allowed to listen to the same sentence more than once. After collecting all the individual scores, the global percentage of correctly identified words was calculated by averaging. In addition, the gain (equivalent SNR increment) over unmodified natural or synthetic speech was calculated based on fits to psychometric functions.

Figure 4 shows the word recognition accuracy scores achieved by our system and also by two baseline systems: unmodified natural speech and artificial speech generated by a reference TTS system. Such a reference TTS is a state-of-the-art HMM-based synthesizer fed on Mel-cepstral coefficients, band aperiodicities and Mel-scaled f_0. Figure 4 also shows the equivalent gain of the proposed system over these two baseline systems. Interestingly, speech generated by the reference TTS is much less intelligible than natural speech in noise regardless of the noise type. This means that the statistical modeling + generation framework has a non negligible impact on speech intelligibility, at least for the amount of training data used in this evaluation. This said, the proposed system clearly outperforms the reference TTS in every noise conditions, the gain being significantly larger when the SNR is low. It is even more intelligible than natural speech in low SNRs, while in high SNRs the proposed enhancing modifications do not totally compensate the loss caused by the statistical synthesis framework. The relative performance of the proposed system with respect to other TTS systems taking part in the Hurricane Challenge is reported in [8].

All these results confirm that it is certainly possible to generate highly intelligible synthetic speech even when the database has been recorded in typical silent conditions and no adaptation data are available. Satisfactory results are obtained just by using only a relatively simple harmonic vocoder and a set of deterministic noise-independent modifications.

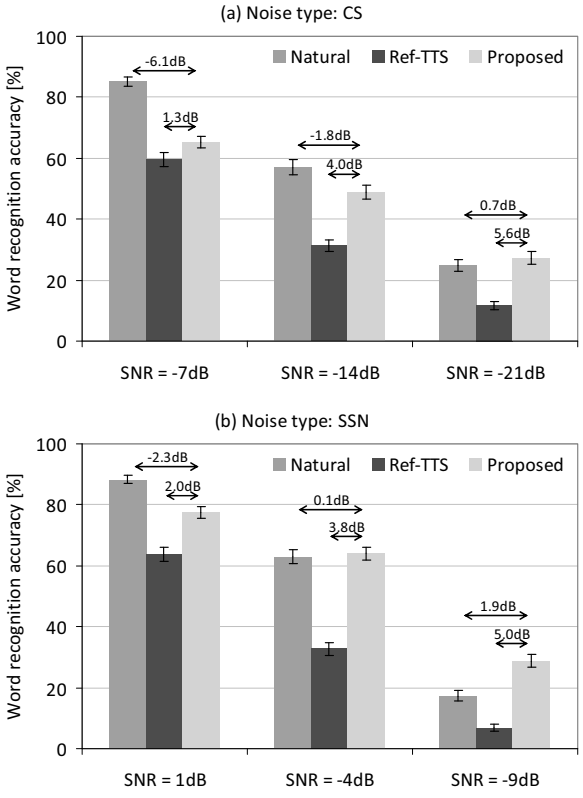

Figure 4: *Results of the evaluation in terms of word recognition accuracy for (a) competing speaker noise and (b) speech-shaped noise. Error bars: ± standard deviation. The numbers above the bars indicate the equivalent gain of the proposed system over the corresponding baseline system.*

5. Conclusions

This paper has described a HMM-based TTS system that includes several noise-independent prosodic and spectral modifications to enhance the intelligibility of synthetic speech in noise. Specifically, modifications are made on duration, pitch level and range, energy contour, formant sharpness, and intensity of particular spectral bands. A relatively simple harmonic model based vocoder provides the necessary flexibility for signal manipulation and facilitates the implementation of the system. The results of a large subjective test confirm the effectiveness of the proposed techniques.

6. Acknowledgements

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03) and the Basque Government (Ber2tek project, IE12-333). Prof. Y. Stylianou held a visiting fellowship from UPV/EHU. We would like to thank Cassia Valentini-Botinhao for providing us with the context labels and question files that were necessary to train the proposed system.

7. References

[1] B. Langner, A.W. Black, "Improving the Understandability of Speech Synthesis by Modeling Speech in Noise", Proc. ICASSP, pp. 265-268, 2005.

[2] B. Picart, T. Drugman, T. Dutoit, "Continuous Control of the Degree of Articulation in HMM-based Speech Synthesis", Proc. Interspeech, pp. 1797-1800, 2011.

[3] T. Raitio, A. Suni, M. Vainio, P. Alku, "Analysis of HMM-Based Lombard Speech Synthesis", Proc. Interspeech, pp. 2781-2784, 2011.

[4] Z.H. Ling, K. Richmond, J. Yamagishi, R.H. Wang, "Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis", IEEE Trans. Audio, Speech & Lang. Process., vol. 17(6), pp. 1171-1185, 2009.

[5] D.Y. Huang, S. Rahardja, E.P. Ong, "Lombard Effect Mimicking", Proc. 7th ISCA Speech Synthesis Workshop, pp. 258-263, 2010.

[6] D. Erro, Y. Stylianou, E. Navas, I. Hernaez, "Implementation of Simple Spectral Techniques to Enhance the Intelligibility of Speech using a Harmonic Model", Proc. Interspeech, 2012.

[7] H. Zen, K. Tokuda, A. W. Black, "Statistical parametric speech synthesis", Speech Commun., vol. 51(11), pp. 1039-1064, 2009.

[8] M. Cooke, C. Mayo, C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge", Proc. Interspeech, 2013.

[9] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Hidden semi-Markov model based speech synthesis", Proc. ICSLP, vol. II, pp. 1397-1400, 2004.

[11] S. Fitt, S. Isard, "Synthesis of regional English using a keyword lexicon", Proc. Eurospeech, pp. 823-826, 1999.

[12] D. Erro, I. Sainz, E. Navas, I. Hernaez, "Improved HNM-based Vocoder for Statistical Synthesizers", Proc. Interspeech, pp. 1809-1812, 2011.

[13] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling", Proc. ICASSP, pp. 229-232, 1999.

[14] D. Erro, I. Sainz, E. Navas, I. Hernaez, "HNM-based MFCC+F0 extractor applied to statistical speech synthesis", Proc. ICASSP, pp. 4728-4731, 2011.

[15] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proc. Institute of Phonetic Sciences, University of Amsterdam, vol. 17, pp. 97-110, 1993.

[16] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification", PhD thesis, ENST, Paris, 1996.
[17] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation", Proc. ICSLP, vol. 3, pp. 1043-1046, 1994.
[18] M.A. Picheny, N.I. Durlach, L.D. Braida, "Speaking clearly for the hard of hearing III: an attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech," J. Speech Hear. Res., vol. 32(3), pp. 600-603, 1989.
[19] M. Koutsogiannaki, M. Pettinato, C. Mayo, V. Kandia, Y. Stylianou, "Can modified casual speech reach the intelligibility of clear speech?", Proc. Interspeech, pp. 579-582, 2012.
[20] J. Villegas, M. Cooke, C. Mayo, "The role of durational changes in the Lombard speech advantage", Proc. Listening Talker Workshop, 2012.
[21] J. C. Krause, "Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement," Ph.D. dissertation, MIT, Cambridge, MA, 2001.
[22] C. Mayo, V. Aubanel, M. Cooke, "Effect of prosodic changes on speech intelligibility", Proc. Interspeech, pp. 1708-1711, 2012.
[23] Y. Lu, M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise", Speech Commun., vol. 51, pp. 1253–1262, 2009.
[24] V. Hazan, A. Simpson, "Cue-enhancement strategies for natural VCV and sentence materials presented in noise", Speech, Hearing and Language, Phonetics and Linguistics, University College London, vol. 9, pp. 43-55, 1996.
[25] B.A. Blesser, "Audio Dynamic Range Compression for Minimum Perceived Distortion", IEEE Trans. Audio & Acoust., vol. 17(1), pp. 22-32, 1969.
[26] T.C. Zorila, V. Kandia, Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression", Proc. Interspeech, pp. 635-638, 2012.
[27] R.J. McAulay, T.F. Quatieri, "Sinusoidal coding", chapter in W.B. Kleijn, K.K. Paliwal (eds.), "Speech coding and synthesis", 1995.
[28] T. Drugman, T. Dutoit, "Glottal-based analysis of the Lombard effect", Proc. Interspeech, pp. 2610-2613, 2010.
[29] E. Godoy, Y. Stylianou, "Unsupervised acoustic analyses of normal and Lombard speech, with spectral envelope transformation to improve intelligibility", Proc. Interspeech, pp. 1472-1475, 2012.
[30] Harvard sentences, from the appendix of "IEEE Recommended Practices for Speech Quality Measurements", IEEE Transactions on Audio and Electroacoustics, vol. 17, pp. 227–246, 1969.