Implementation of Simple Spectral Techniques to Enhance the Intelligibility of Speech using a Harmonic Model

Daniel Erro[1], *Yannis Stylianou*[2,1], *Eva Navas*[1] *and Inma Hernaez*[1]

[1]Aholab Signal Processing Laboratory, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2] Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
{derro,eva,inma}@aholab.ehu.es, yannis@csd.uoc.gr

Abstract

We have designed a system that increases the intelligibility of speech signals in noise by manipulating the parameters of a harmonic speech model. The system performs the transformation in two steps: in the first step, it modifies the spectral slope, which is closely related to the vocal effort; in the second step, it amplifies low-energy parts of the signal using dynamic range compression techniques. Objective and subjective measures involving speech-shaped noise confirm the effectiveness of these simple methods. As the harmonic model has been used in previous works to implement the waveform generation module of high-quality statistical synthesizers, the system presented here can provide the synthesis engine with a higher degree of control on the intelligibility of the resulting artificial speech.

Index Terms: speech intelligibility in noise, harmonic model, speech synthesis, spectral tilt, dynamic range compression.

1. Introduction

Speech synthesizers are usually trained from clean speech databases recorded by professional speakers in silent environments. Consequently, when synthetic speech is played in noisy conditions it is often hard for listeners to understand the message. For speech synthesizers to be more practical in different contexts, it is desirable to provide some kind of control over the voice characteristics that play a crucial role in intelligibility.

There are basically two ways of modifying the synthetic speech to make it more intelligible in noisy conditions: (i) recording a new database in the desired conditions and then using voice conversion [1], speaker adaptation [2][3], inter/extrapolation [3], or any statistical mapping technique [4]; (ii) using the original clean speech database and then applying expert knowledge and signal processing techniques to enhance the output of the synthesizer [5][3]. Although the latter strategy may not reproduce the real behaviour of the speakers in noisy environments as accurately as the former, it has the advantage that it is costless: it avoids the need of recording new databases or retraining the underlying models. The solutions that belong to this second category can be classified into two groups: parametric and non-parametric approaches. Non-parametric approaches do not consider any specific speech model and they may operate in the time and/or the frequency domain. In contrast to this, parametric approaches require the parameters of a specific speech model to be estimated from the signal before modification. In the case where the system input is natural speech, non-parametric approaches are typically more adequate because of their inherent efficiency and quality. In the framework of statistical parametric speech synthesis [6], however, parametric approaches present some interesting properties. In these systems, given an input text, the synthesis engine generates a set of vectors containing a parametric representation of signals. These parameters are interpreted and processed by vocoders to get the output speech waveform. In this framework, the use of parametric approaches does not imply significant computational cost if the speech enhancement procedures are defined in the same parametric domain as the waveform generation process.

Speech models handling an explicit parameterization of the vocal tract and the glottal source [3][7] are very flexible and provide a high degree of control over voice quality and phonation type. Nevertheless, the robustness of the analysis techniques involved in these models is still an open issue [8]. In practice, the vast majority of the statistical synthesizers today operate on less sophisticated parametric representations of speech such as Mel-cepstral coefficients. For this reason, speech enhancement procedures that are compatible with these parameters are highly desirable.

It was shown in [9][10] that a harmonic model (HM) can be effectively applied to reconstruct speech from a sequence of Mel-cepstral vectors generated by a statistical synthesizer. In this work, we show that the parameters of the same HM can be manipulated to increase speech intelligibility in noise, even without modifying the energy of the signal. We propose a two-step transformation. During the first step, the spectral slope is increased to mimick the effect of higher vocal effort. During the second step, the energy of the signal is redistributed over time to amplify meaningful low-energy parts of the signal. This transformation operates basically on the harmonic amplitudes and can be easily integrated into an HM-based waveform reconstruction module of a statistical synthesizer, providing a higher degree of control on intelligibility at the expense of an almost negligible increment of the computational load.

The rest of the paper is structured as follows. Section 2 describes the frequency slope modification and the dynamic range compression procedure in the context of HM. In section 3, objective and subjective measures are used to configure and evaluate the performance of the proposed system. Finally, the conclusions of this research work are summarized in section 4.

2. Description

The HM assumes that locally stationary segments of a given speech signal can be decomposed into a series of harmonically related sinusoids represented by their frequencies, amplitudes and phases:

$$s[n] \cong \sum_{i=1}^{I} A_i \cos(i\omega_0 n + \varphi_i) \ , \quad \omega_0 = 2\pi f_0 / f_s \qquad (1)$$

where f_s is the sampling frequency and I is the number of harmonics inside the analysis band (0, f_s/2). Assuming an adequate value of f_0, typically 100Hz, this model gives good perceptual results even in unvoiced speech segments.

In natural signals, the parameters of HM can be extracted by analyzing short frames at 10ms (or lower) frame shift via least squares analysis [11] or spectral peak picking algorithms [12]. In synthetic signals generated by a statistical engine, vectors containing Mel-cepstral coefficients (the most common spectral envelope representation in synthesis tasks) can be translated into harmonic parameters to perform high-quality waveform reconstruction [9][10]. Therefore, transformations defined in the HM domain are useful in both scenarios.

The proposed system operates entirely on the parameters of HM, more specifically on the amplitudes (phase modifications are beyond the scope of this paper). It consists of two transformation steps to be applied in cascade. These two transformations are described in the next subsections.

2.1. Spectral slope modification

In general, Lombard speech is characterized by a higher vocal effort, which has an impact on the spectral tilt [13][14][15]. We propose a simple spectral transformation that modifies the harmonic amplitudes by a constant slope measured in dB/decade and then multiplies them by a normalization term to preserve the energy at frame-level. A similar approach was followed in [16] to modify the emotional content of speech signals. This is done frame-by-frame as follows:

$$A'_i = A_i \cdot i^{m/20} \qquad (2)$$

where m is the spectral slope measured in dB/decade and i denotes the harmonic order. In order to preserve the original energy of the frame we multiply all the modified amplitudes by a correction factor. According to Parseval's theorem, this factor can be estimated directly from the harmonic amplitudes:

$$A''_i = A'_i \cdot \sqrt{\frac{\sum_{i=1}^{I} A_i^2}{\sum_{i=1}^{I} A_i'^2}} \qquad (3)$$

Please note that in (2) and (3) the frame index k has been omitted for clarity. Positive values of m enhance the higher frequencies and therefore the intelligibility. Nevertheless, as the energy is kept constant after the spectral modification, there is a slight energy decrement at lower frequencies. For this reason, m cannot be given arbitrarily high values. Several factors have to be considered when trying to determine the optimal value of m: the perceptual quality of the enhanced signal, the specific noise level, and the spectral tilt of the original voice. In the sort of voices that are typically used in speech synthesizers, m ranges from 15 to 30dB/decade for negative signal-to-noise ratio (SNR).

2.2. Dynamic range compression

Despite the usefulness of the spectral slope modification, some specific consonants are likely to be masked by noise due to their low energy. Most of these phonemes play a decisive role in intelligibility: plosives, fricatives, vocalic onsets and offsets, and

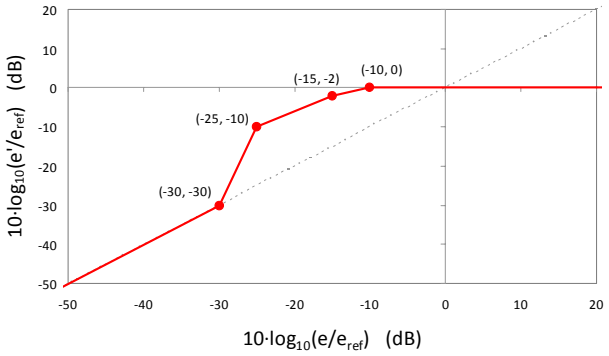

Figure 1: *Nonlinear energy mapping curve for DRC.*

nasals. In [17], these acoustic-phonetic events in the speech signal are targeted for selective enhancement. Particularly it has been shown that selective reinforcement of bursts and vocalic onsets can provide significant improvements to the intelligibility of the subsequently degraded speech signal, even for the same overall signal-to-noise ratio [17]. Meanwhile, other phonemes, mainly vowels, are likely to carry more energy than required by human listeners to decode the message. One possible solution for this is redistributing the energy of the signal in such manner that significant low-energy portions are amplified. Dynamic Range Compression (DRC) techniques [18][19] are typically used for this exact purpose. We propose a DRC implementation that involves only the harmonic amplitudes. First, we compute the local energy of the signal as

$$e^{(k)} = \sum_{i=1}^{I^{(k)}} A_i^{(k)2} \qquad (4)$$

where k denotes the frame index and $I^{(k)}$ is the local number of harmonics, which depends on the local f_0. Note that this local energy was preserved during the spectral slope modification step. Next, we apply a nonlinear mapping function to $e^{(k)}$:

$$e'^{(k)} = f_{DRC}\left\{e^{(k)}\right\} \qquad (5)$$

Figure 1 shows the input-output energy function we adopted for this work. As it can be seen, f_{DRC} is a piecewise linear function in the dB scale that is defined with respect to a reference level, e_{ref}, which is somehow related with the energy range of the original signal. In this work,

$$e_{\text{ref}} = 0.3 \cdot \max\left\{e^{(1)} \ldots e^{(K)}\right\} \qquad (6)$$

Thus, the energy at insignificant parts of the signal (below -30dB in Figure 1, which corresponds mainly to silences, pauses, etc.) is kept unaltered, while energies at intermediate levels are moved towards the maximum. As a consequence of this, low-energy phonemes are amplified. As the total energy of the signal is altered by this nonlinear energy mapping, we renormalize it as follows:

$$e''^{(k)} = e'^{(k)} \cdot \frac{\sum_{k=1}^{K} e^{(k)}}{\sum_{k=1}^{K} e'^{(k)}} \qquad (7)$$

Finally, this new energy contour is imposed to the local harmonic amplitudes via local multiplicative factors:

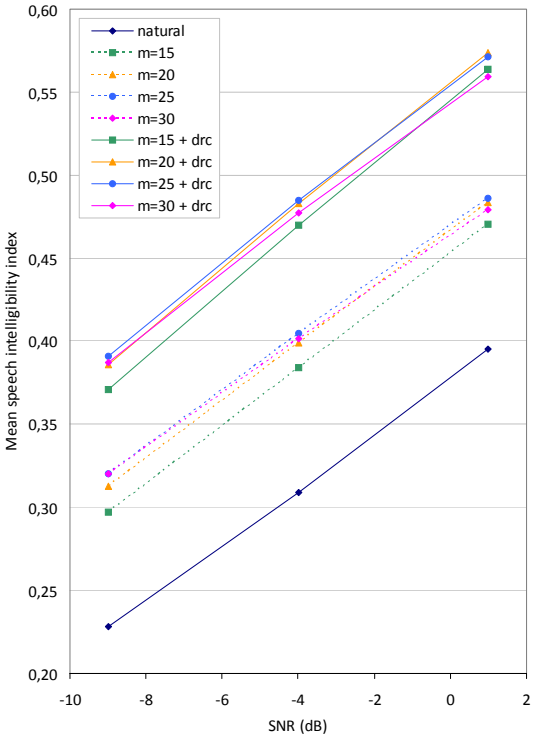

Figure 2: *Objective SII scores for speech-shaped noise.*

$$A_i'''^{(k)} = A_i''^{(k)} \cdot \sqrt{\frac{e''^{(k)}}{e^{(k)}}} \qquad (8)$$

where A_i'' are the amplitudes that resulted from (3) during the previous step.

The output speech is generated from the amplitudes given by (8), while the frequencies and phases are kept unaltered. In our implementation, speech reconstruction is carried out by generating short frames with constant harmonic parameters and combining them by overlap-add (OLA). Since the described procedures involve a relatively low number of operations, the inclusion of the modification module does not imply a significant increment of the computational load.

3. Experiments and Discussion

The database used in both objective and subjective experiments contained 240 phonetically balanced sentences from the Harvard corpus [20]. All the sentences were recorded by a male native British English talker and then paired with masking noises at three different SNR values: –9dB, –4dB and 1dB. Speech-shaped noise was used instead of white noise. In other words, the spectrum of the noise was shaped to match the average spectrum of human speech in order to provide a more realistic evaluation scenario. The sampling rate for all files was 16kHz.

A first experiment based on the Speech Intelligibility Index (SII) was conducted to objectively assess the performance of the enhancement system. For the computation of SII we followed the steps described in [21], towards what is referred to as Extended SII. First, a FIR filter bank is used to filter speech and noise signals into 21 critical bands [22], using linear FIR filters of type

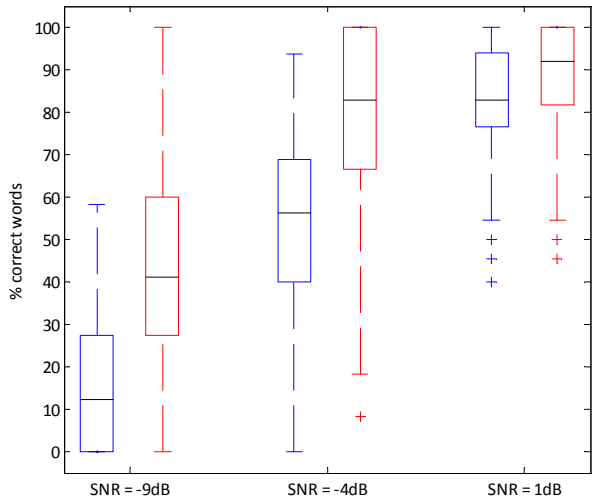

Figure 3: *Results of the subjective test for speech-shaped noise (median, quartiles and extreme values of the distributions). Blue: natural signals. Red: enhanced signals.*

I and order 200. Next, the time varying intensity of the signal is computed for each output of the filter bank. For this, non overlapped rectangular windows are used with window lengths ranging from 35ms at the lowest band (center frequency 50Hz) to 10ms at the highest band (center frequency 7kHz). The windows are aligned such that they end simultaneously [21]. The intensity level is normalized to dBSPL using the absolute threshold of hearing (10^{-12} watts). At a given instant, the instantaneous SII is computed following a standard procedure (ANSI S3.5-1997, [23]) using the so-called speech perception in noise (SPIN) weighting function and the estimated speech and noise normalized intensities. Finally, the SII for a speech-in-noise condition is determined by simple averaging across all the instantaneous SII values. The objective intelligibility score computed by SII was validated using results from a listening test described in [24], where 88 native English speakers had to listen to 96 different sentences in noise. Sentences were generated by an HMM-based synthesizer. Four different types of noise were used (speech-shaped, cafeteria, car and high-frequency noise) at four levels of SNR which shown to roughly correspond, for each of the previously mentioned noise types, to word prediction probabilities of 0.2, 0.4, 0.6, and 0.8. It is worth to mention that in [24] it was reported that the SII objective measure was poorly correlated (compared to other objective measures) with the subjective results (average correlation about 0.5). Using the same data, our implementation of SII gives a much higher correlation score (0.8), which is comparable with the best predictors listed in [24]. Therefore, our objective measure can be considered a good predictor of speech intelligibility, at least under the noise conditions tested in [24].

Figure 2 shows the objective scores achieved by the system for four different values of m (15, 20, 25, and 30dB/dec.), with and without DRC. According to these scores, the two suggested modifications prove useful in terms of SII, each yielding an increment of approximately 0.1 points in the SII scale. Interestingly, within the interval from 20 to 25dB/dec., the choice of m has little impact on the objective scores, although a slight dependence is observed between the optimal value of m and the specific noise level. Theoretically it is possible to

develop an automatic method based on the same principles as the SII to estimate the optimal *m* for a given input noise. However, the design of such a method is not straightforward because it should consider not only the spectral properties of the modified signal but also the perceptual quality, which has a direct impact on the probability that the message is correctly understood by human listeners.

A perceptual test was carried out at CSTR, Univ. of Edinburgh, in which 78 participants with British English as native language listened to original and modified utterances mixed with speech-shaped noise in sound-isolated booths and were asked to type what they heard. The first 180 sentences of the database described at the beginning of this section were used for the test. The listeners were not allowed to hear the same sentence more than once. With regard to the configuration of the system, the value of *m* was adjusted according to the results of the previous objective test: for each SNR, the optimal *m* was determined by parabolic fitting around the maximum of the SII(*m*) curve (see Figure 2). The results of this test are shown in Figure 3, where the distribution of the percentage of correct words are plotted for natural (blue color) and enhanced signals (red color). Such distributions were calculated from the scores given to all the test sentences by all the listeners. The subjective results confirm that, despite its simplicity, the proposed system increases the number of correct words substantially with respect to the original natural voice. The improvements are visible mainly in low SNR conditions, where the number of correct words increases by factor 4. Similar informal listening tests revealed that the system performed equally well for other voices and languages. In these tests, *m* was set to 20-25.

4. Conclusions

We have proposed a system to enhance the intelligibility of speech in noise without increasing the energy of the input signal by manipulating the parameters of a harmonic model. The system consists of two simple modification steps: tilt modification and compression of the energy range. Objective and subjective tests confirm the effectiveness of the proposed system in the presence of speech-shaped noise. These findings will enable the development of noise-adaptive speech synthesis engines based on a harmonic model, which are used in some state-of-the-art systems to produce high-quality synthetic speech from conventional Mel-cepstral vector sequences. Experiments involving other types of noise are in progress. Future works will aim at exploring phase modifications.

5. Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation (BUCEADOR Project, TEC2009-14094-C04-02) and by the European Commission (LISTA Project, FET-Open grant number 256230). We would like to thank Cassie Mayo, Vasilis Karaiskos and Martin Cooke for their contribution in the subjective evaluation of the system, and Cassia Valentini-Botinhao for providing us with the material used in the perceptual test in [24].

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

[7] P. Lanchantin, G. Degottex, X. Rodet, “A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method”, Proc. ICASSP, pp. 4630-4633, 2010.

[8] T. Drugman, B. Bozkurt, T. Dutoit, “A Comparative Study of Glottal Source Estimation Techniques”, Computer Speech & Language, vol. 26(1), pp. 20-34, 2012.

[9] D. Erro, I. Sainz, E. Navas, I. Hernaez, “HNM-based MFCC+F0 Extraxtor applied to Statistical Speech Synthesis”, Proc. ICASSP, pp. 4728-4731, 2011.

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

[11] Y. Stylianou, “Modeling Speech based on Harmonic plus Noise Models”, Lecture Notes in Computer Science, 2005.

[12] R.J. McAulay, T.F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation”, IEEE. Trans. Acoust., Speech, and Sig. Process., vol. 34(4), pp. 744-754, 1986.

[13] J.C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” J. Acoust. Soc. Am., vol. 93(1), pp. 510-524, 1993.

[14] 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.

[15] T. Drugman, T. Dutoit, “Glottal-based Analysis of the Lombard Effect”, Proc. Interspeech, 2010.

[16] D. Erro, E. Navas, I. Hernaez, I. Saratxaga, “Emotion Conversion based on Prosodic Unit Selection”, IEEE Trans. Audio, Speech, & Lang. Process., vol. 18(5), pp. 974-983, 2010.

[17] 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.

[18] B. A. Blesser, “Audio Dynamic Range Compression for Minimum Perceived Distortion”, IEEE Trans. Audio & Acoust., vol. 17(1), 1969.

[19] J. Birch, “Evaluation of Clipping and Limiting Amplifiers”, Project Report no. 636, Audio Branch, Engineering Division IBS/ET, United States Information Agency, 1974.

[20] 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.

[21] K.S. Rhebergen, N.J. Versfeld, “A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners”, J. Acoust. Soc. Am., vol. 117(4), pp. 2181-2192, 2005.

[22] B. Truax (editor), “Handbook for Acoustic Ecology”, 2nd Ed. Simon Fraser University and ARC Publications, 1999.

[23] ANSI S3.5-1997, “American national standard methods for calculation of the speech intelligibility index”, American National Standards Institute, New York, 1997.

[24] C. Valentini-Botinhao, J. Yamagishi, S. King, “Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?”, Proc. Interspech, 2011.