

# Time and Frequency Dependent Amplification for Speech Intelligibility Enhancement in Noisy Environments

Henk Brouckxon<sup>1</sup>, Werner Verhelst<sup>1</sup>, Bart De Schuymer<sup>2</sup>

<sup>1</sup>Vrije Universiteit Brussel, dept. ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium

<sup>2</sup>TELEVIC nv, Corporate R&D Department, Leo Bekaertlaan 1, B-8870 Izegem, Belgium

hbrouckx, wverhels@etro.vub.ac.be, b.deschuymer@televic.com

## Abstract

When speech is presented through loudspeakers in a noisy environment, the background noise can significantly decrease speech intelligibility. Because the amplitude and spectrum of the background noise can vary over time (and because high loudness levels are to be avoided for listener comfort), choosing proper speech equalization and master gain settings for a public address system can be a difficult task. In this paper, we propose an adaptive digital signal processing algorithm that applies a frequency and time dependent gain strategy to the speech signal in order to enhance its intelligibility in noise with a minimal increase of the overall sound energy level. An alternative version of the system can also be used to maximise speech intelligibility without increasing the overall energy level of the signal. The proposed algorithm makes use of the psycho-acoustic masking properties of the human hearing system and relies on the importance of the formant information for speech intelligibility.

**Index Terms:** speech intelligibility, psycho-acoustics, noise masking

## 1. Introduction

In many applications, such as public addressing systems, speech messages are presented in acoustic environments with a significant amount of background noise. The amount and type of this background noise can vary strongly over time and largely determines the intelligibility of the messages, especially at times when the noise level is large relative to the speech presentation level. This can for example be experienced in train stations, where babble noise from the passengers and sounds from arriving and departing trains are typically present. Even though it is technically possible to amplify the speech sufficiently to remain audible above the noise in many cases, a tradeoff has to be made because too large levels of speech are uncomfortable for the listeners, and can even be detrimental to intelligibility. Some state-of-the-art systems partially overcome this problem by adaptively increasing the overall speech amplification when noise levels get high. In this way, a predetermined minimum signal-to-noise ratio ( $SNR$ ) can in most cases be maintained [1], improving intelligibility over a large range of noise levels. In this paper, we propose a system for time and frequency dependent speech amplification that ensures intelligibility of the speech at lower overall signal levels than the  $SNR$ -based systems, and that can also be used to optimise intelligibility at a given signal level or  $SNR$ . This system determines the necessary amplification factors for intelligibility enhancement, based on a psycho-acoustic intelligibility criterion.

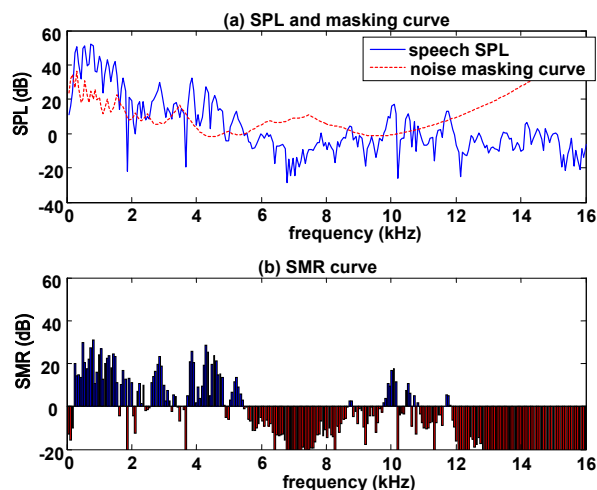


Figure 1: Illustration of the results of a psycho-acoustic masking model for audibility evaluation

## 2. Speech intelligibility estimation

Although objective measures like the Speech Transmission Index (STI) [2] exist that predict the influence of noise and signal distortions on the intelligibility of speech in general, little or no objective measures are available that evaluate their influence on the intelligibility of a specific utterance. For our system, we therefore developed an intelligibility evaluation criterion, based on the audibility of the speech in the frequency regions that are most important for speech intelligibility.

### 2.1. Audibility: psycho-acoustic masking

Psycho-acoustic research has shown that a threshold Sound Pressure Level ( $SPL$ ) exists for each frequency, below which the human hearing system does not perceive any sound [3, 4]. This hearing threshold is increased when another sound (e.g. background noise<sup>1</sup>) is already present in the listening environment. In this way, the background noise signal 'masks' parts of

<sup>1</sup>In this text, we assume that the background noise is known. In many cases, a noise reference can easily be obtained by placing measurement microphones in strategic locations of the presentation environment. In case the microphones need to be placed such that they will also pick up the loudspeaker signals, adaptive interference cancellers can be used to remove the loudspeaker signals from the microphone signal [1, 5].

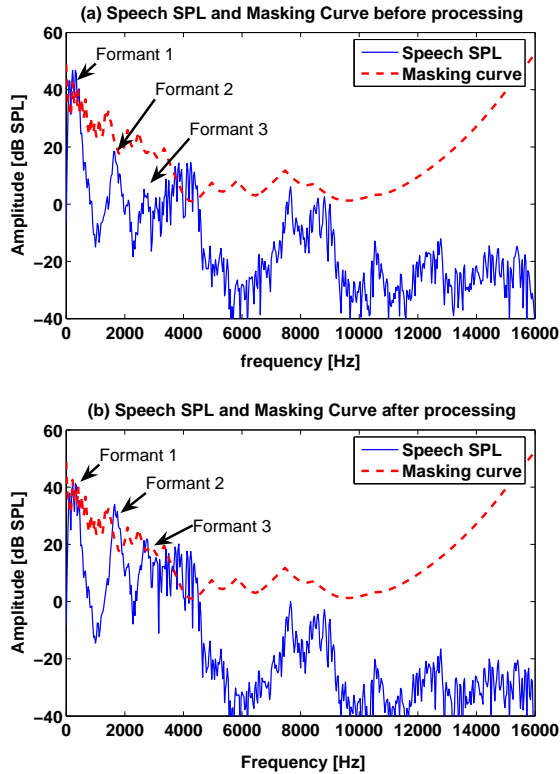


Figure 2: Speech SPL before and after processing

the spectrum of the presented speech signal, and makes it difficult or impossible to comprehend. Psycho-acoustic masking models have been developed that can predict the hearing threshold for an average listener, based on the instantaneous background noise. These models are in general use for audio coding and compression where they predict the amount of quantization noise that can be masked by the audio (e.g. used in MP3, WMA and AAC encoders) [6]. In Figure 1.a an example is given of the typical characteristics of a psycho-acoustic masking model: the instantaneous spectral energy content of the speech (represented by the Speech Sound Pressure Level (*speech SPL*) curve) and the *hearing threshold* curve, corresponding to the background noise. Based on these curves, the Signal-to-Masking threshold Ratio,  $SMR = \text{speech SPL} - \text{hearing threshold}$  (Figure 1.b), expresses which frequency components of the presented speech remain audible in the noise background. Frequency components with an  $SMR$  below 0 dB will not be discernible above the noise for average listeners, while frequency components with a positive  $SMR$  remain audible, with higher  $SMR$  values corresponding to better audibility. While the psycho-acoustic model predicts the audibility of the individual frequency components of the speech signal, inaudibility of some components does not necessarily imply unintelligible speech because various frequency components have a different impact on speech intelligibility. Since the frequency regions around the first three formants (Figure 2) are most important for speech intelligibility [7] we will use the audibility ( $SMR$  value) of the first three formant regions as an operational criterion for the intelligibility of the speech signal.

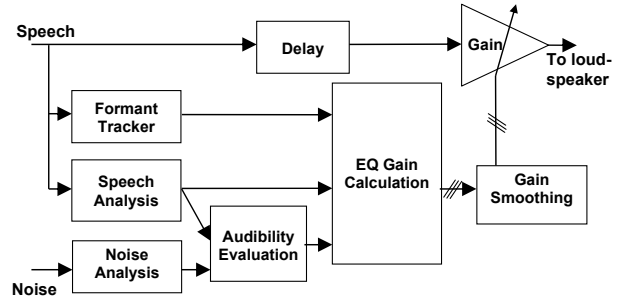


Figure 3: Block diagram for the time and frequency dependent gain system

### 3. Amplification strategy

From section 2 one could conclude that, in order to ensure speech intelligibility, the *Speech* signal must be amplified such that its first three formants always stay above the hearing threshold ( $SMR > 0\text{dB}$ ). However, since the  $SMR$  can be different for each formant, it is not desirable to amplify the whole signal with a same large constant factor: with a time and frequency dependent gain control, intelligibility can be ensured with a minimal increase in overall loudness. In this section, we present an efficient time and frequency dependent amplification strategy (Figure 3) for intelligibility enhancement.

#### 3.1. Step 1: formant and $SMR$ calculation

The *Formant Tracker* block in Figure 3 performs voicetype classification and tracks the evolution of the formants in the incoming *Speech* signal. The speech is segmented into frames of length 10 ms, for which a voicetype (voiced, unvoiced, fricative or pause) is determined using a computationally simple method similar to [8, 9], based on frame energy, zero-crossing frequency and spectral energy distribution. Formant tracking is performed on the same frames of length 10 ms based on an LPC pole-tracking algorithm, similar to [10, 11]. An example of a speech spectrogram and its corresponding formant tracks is shown in Figure 5.

The *Speech Analysis*, *Noise Analysis* and *Audibility Evaluation* blocks form a psycho-acoustic system that determines the  $SMR$  of the *Speech* signal when presented in the noisy acoustic environment (*Noise*). The *Noise Analysis* block determines the hearing threshold corresponding to *noise*, and the *Speech Analysis* block transforms the incoming *speech* signal into its SPL characteristic. The *Audibility Evaluation* block then combines these into the  $SMR$  characteristic. In our current implementation, these blocks use the MPEG-1 layer II psycho-acoustic model 1 [6].

#### 3.2. Step 2: raw gain calculation (*EQ gain calculation block*)

##### 3.2.1. Voiced and unvoiced sounds :

For each formant trajectory  $F_i(n)$  ( $i = 1 \dots 3$ ), the maximum  $SPL$  value  $SPL_i(n)$  and maximum hearing threshold value  $THR_i(n)$  are determined in frequency regions  $R_i(n)$  around the  $F_i(n)$  that correspond to the average formant widths (Eq. 1 and 2). The corresponding formant  $SMR$  value  $SMR_i(n)$  is then calculated as given in Eq. 3, and used to calculate a raw time-dependent gain trajectory  $G_{i,raw}(n)$ , corresponding to a

predefined goal  $SMR$  value  $SMR_{i,goal}$  (Eq. 4). For these sounds, the overall gain factor  $G_{0,raw}(n)$  was set to 0 dB.

$$SPL_i(n) = \max(SPL(n, R_i(n))) \quad (1)$$

$$THR_i(n) = \max(THR(n, R_i(n))) \quad (2)$$

$$SMR_i(n) = SPL_i(n) - THR_i(n) \quad (3)$$

$$G_{i,raw}(n) = \max(0, SMR_{i,goal} - SMR_i(n)) \quad (4)$$

with

$$R_1(n) = [F_1(n) - 100Hz, F_1(n) + 100Hz]$$

$$R_2(n) = [F_2(n) - 200Hz, F_2(n) + 200Hz]$$

$$R_3(n) = [F_3(n) - 500Hz, F_3(n) + 500Hz]$$

### 3.2.2. Fricative sounds and pauses:

Because the spectral envelope of fricative sounds typically corresponds to a highpass characteristic, here only a single 'formant' region is considered that spans the frequency range  $R(n) = [4000 Hz, 8000 Hz]$ . The raw overall gain trajectory  $G_{0,raw}(n)$  is determined as the gain that increases the average  $SMR$  in this region to a predefined goal  $SMR_{f,goal}$ . For fricative sounds,  $G_{1,raw}(n)$ ,  $G_{2,raw}(n)$  and  $G_{3,raw}(n)$  are set to zero dB.

$$G_{0,raw}(n) = \max(0, SMR_{f,goal} - \text{mean}(SMR(n, R(n)))) \quad (5)$$

During pauses, the raw overall gain trajectory and the raw formant gain trajectories are maintained at the values corresponding to the last frame before the pause.

### 3.3. Step 3: gain smoothing

Since the raw gain values from Eq. 4 and 5 can vary strongly in between the formants, and for each formant can change abruptly over time, a direct application of the gains to the speech signal would typically result in audible distortion of the signal. To avoid this, the gain values are smoothed in time and frequency (*Gain Smoothing* block).

In a first phase, the changes in the spectral tilt of the speech signal are limited by imposing boundaries for the instantaneous differences between the raw formant gains  $G_{i,raw}(n)$ . Because the first formant  $F_1(n)$  typically contains most energy and its gain trajectory is most stable,  $G_{1,raw}(n)$  is chosen as a reference  $G_1^*(n)$ , and the gain trajectories for formants  $F_2(n)$  and  $F_3(n)$  are bounded within a region, defined by  $\alpha$ , the maximum change in spectral tilt expressed in dB/Hz, as shown in Eq. 6.

$$\begin{aligned} G_i^*(n) &= G_{i,min}(n) \text{ if } G_{i,raw}(n) < G_{i,min}(n) \\ &= G_{i,max}(n) \text{ if } G_{i,raw}(n) > G_{i,max}(n) \\ &= G_{i,raw}(n) \text{ otherwise} \end{aligned} \quad (6)$$

$$\begin{aligned} \text{with } G_{i,max}(n) &= G_1^*(n) + (\alpha * (F_i(n) - F_1(n))) \\ G_{i,min}(n) &= G_1^*(n) - (\alpha * (F_i(n) - F_1(n))) \end{aligned}$$

The value of  $\alpha$  is a tradeoff between the audibility of the speech modifications (if  $\alpha$  is large, a time-varying highpass effect is sometimes perceived) and the intelligibility enhancement. For low noise levels, a small  $\alpha$  value is typically preferred, because this guarantees a natural perception of the speech. When very high noise levels are expected however,  $\alpha$  should be increased to improve the intelligibility-energy ratio of the speech

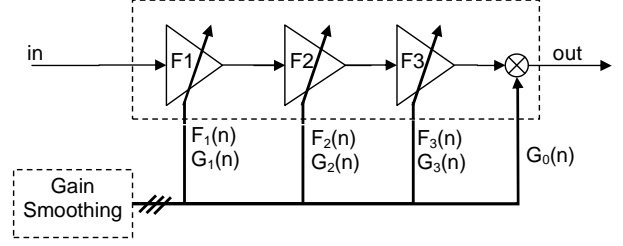


Figure 4: Block diagram for the *Gain* subsystem

signal, which is in this case more important. By increasing alpha with the noise level, an adaptive system can thus be constructed that applies a time-dependent gain for low noise levels (natural perception), and adds spectral intelligibility optimisation when noise levels get high.

In the second smoothing phase, all formant and overall gains are smoothed over time using a first-order lowpass filter, as in Eq. 7 and 8. Also, the overall gain  $G_0(n)$  is taken into account in the formant gains (Eq. 9).

$$G_0(n) = \beta_0 G_0(n-1) + (1 - \beta_0) G_{0,raw}(n) \quad (7)$$

$$G_i^{**}(n) = \max(0, (\beta_i G_i^{**}(n-1) + (1 - \beta_i) G_i^*(n))) \quad (8)$$

$$G_i(n) = G_i^{**}(n) - G_0(n) \quad (9)$$

### 3.4. Applying the gain (*Gain* block)

The time- and frequency dependent gain factors  $G_i(n)$  ( $i = 0 \dots 3$ ) are applied to the speech signal as shown in Fig. 4. Three second-order parametric equaliser sections [12] (blocks  $F1$ ,  $F2$  and  $F3$ ) are tuned to the formant tracks, and apply the corresponding formant gain factors  $G_1(n)$ ,  $G_2(n)$  and  $G_3(n)$ . Analogous to the processing in the intelligibility analysis phase, the 3dB bandwidths of these equaliser sections are set to correspond to the average bandwidths of the formant regions (200 Hz for  $F1$ , 500 Hz for  $F2$  and 1000 Hz for  $F3$ ).  $G_0(n)$  is implemented as an overall, frequency independent gain factor.

## 4. Evaluation and future work

A first subjective evaluation of the system was performed, based on the Dutch (Flemish) LIST intelligibility test [13]. Five persons (male, 24-27 years of age) with normal hearing participated in this test, all native Dutch (Flemish) speakers. In this test, the listener's Speech Reception Thresholds ( $SRT$ )<sup>2</sup> for unprocessed speech and for speech that was processed by our enhancement algorithm were measured using lists of 10 prerecorded utterances from [13]. For each list, an initial signal-to-noise ratio ( $SNR$ ) was determined by repeating the first utterance at gradually increasing  $SNR$ s (2 dB  $SNR$  increase per repetition, starting at an unintelligible -12 dB  $SNR$  level) until a number of predefined keywords were correctly understood. Utterances 2 through 10 were each presented only once, at an  $SNR$  level that is adapted using the following rules:

- if all keywords in sentence  $i$  are correctly understood, the  $SNR$  is decreased by 2 dB for sentence  $i + 1$

<sup>2</sup>the signal-to-noise ratio at which 50% of a given list of utterances are intelligible

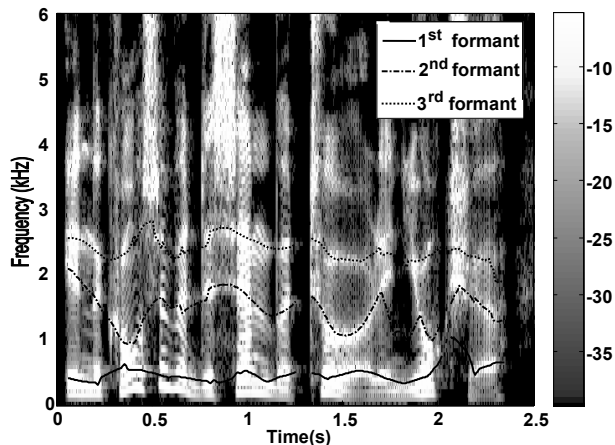


Figure 5: Spectrogram and corresponding formant tracks 1-3. White areas in the spectrogram correspond to high energy levels

- If not all keywords in sentence  $i$  are correctly understood, the  $SNR$  is increased by 2 dB for sentence  $i + 1$

The  $SRT$  was calculated as the average  $SNR$  level for utterances 2 through 10. All utterances were presented in a babble noise background, at a fixed and comfortable noise level (chosen prior to the tests by the listeners), using a good-quality pc audio interface and Sennheiser HD555 headphones. For all test persons, the  $SRT$  of the processed utterances was 3.5 to 4.5 dB lower than the  $SRT$  for unprocessed utterances. This result indicates that our processed speech can retain its intelligibility in an environment with an  $SNR$  that is up to 4 dB worse than for normal, unprocessed speech.

## 5. CONCLUSION

In this paper, a system for intelligibility enhancement of speech presentation in a noisy environment was proposed. Based on the psycho-acoustic properties of the human hearing system, this system applies a time and frequency dependent gain to the speech signal in order to make its first three formants audible with as little amplification as possible. A formal evaluation of this system indicated an intelligibility enhancement of approximately 4 dB  $SRT$ .

## 6. ACKNOWLEDGEMENTS

This research was performed with the support of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Vlaanderen), projects IWT040803(SMS4PA-II) and IWT070317(CoMiCS).

## 7. References

- [1] G. Rombouts, T. van Waterschoot, K. Struyve, and M. Moonen, "Acoustic feedback suppression for long acoustic paths using a nonstationary source model," *IEEE Trans. on Signal Processing*, vol. 54, no. 9, September 2006.
- [2] H. J. M. Steeneken and T. Houtgast, "Phoneme-group specific octave-band weights in predicting speech intelligibility," *Speech Communication*, vol. 38, pp. 399–411, 2002.
- [3] H. Fletcher and W. Munson, "Loudness, its definition, measurement and calculation," *Journal of the Acoustical Society of America*, vol. 5, pp. 82–108, 1933.
- [4] "ISO/IEC standard 226:2003, Acoustics - Normal equal-loudness-level contours," 2003.
- [5] C. F. N. Cowan and P. M. Grant, *Adaptive Filters*. Prentice Hall, 1985.
- [6] "ISO/IEC standard 11172-3, Information technology - Coding of moving pictures and associated audio for digital storage media at about 1.5 mbit/s part 3: Audio," 1993.
- [7] G. Fant, *Acoustic theory of speech production*. Mouton, 's Gravenhage, Nederland, 1960.
- [8] W. Mattheyses, W. Verhelst, and P. Verhoeve, "Robust pitch marking for prosodic modification of speech using TD-PSOLA," in *proc. of the second annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2006)*, march 2006, pp. 43–46.
- [9] M. Greenwood and A. Kinghorn, *Saving: Automatic silence/unvoiced/voiced classification of speech*. Department of Computer Science, The University of Sheffield, 1999.
- [10] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE trans. on Acoustics, Speech and Signal Processing*, April 1974.
- [11] R. Schafer and L. Rabiner, "System for automatic formant analysis of voiced speech," *Journal of the Acoustical Society of America*, vol. 57, no. 2, pp. 634–648, 1970.
- [12] P. Regalia and S. Mitra, "Tunable digital frequency response equalisation filters," *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol. ASSP-35, no. 1, pp. 118–120, 1987.
- [13] A. van Wieringen and J. Wouters, "LIST en LINT: Nederlandstalige Spraak-audiometrielijsten met Zinnen en Getallen," 2005.