

MUSIC INJECTION FOR SUBJECTIVE SPEECH ENHANCEMENT AND THE PSYCHOACOUSTIC PLEASANTNESS ANALYSIS

Hua Bao, Issa M.S. Panahi, Philipos C. Loizou, Yi Hu

Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas, 75080
Email: hua.bao@student.utdallas.edu

ABSTRACT

In this paper we propose a method for subjective speech enhancement. Traditional noise reduction methods, although powerful in improving the SNR of speech signal, have the drawback of generating unpleasant residual noise, which makes the processed speech annoying. Music injection is adopted to improve the speech quality with respect to pleasantness. A prediction model of pleasantness based on four psychoacoustic measures (loudness, sharpness, roughness and tonality) is used to evaluate the proposed method. Three different types of instrumental music are taken to analyze the performance of our method. Subjective tests are also conducted to verify the objective pleasantness model.

Index Terms— Subjective speech enhancement, music injection, spectral subtraction, psychoacoustic pleasantness

1. INTRODUCTION

In recent years, the enhancement for noisy speech has gained an increasing interest. Various speech enhancement algorithms have been proposed to reduce noise [1]. They can be categorized as four different types: spectral subtractive, subspace, statistical-model based and Wiener algorithms. Although these methods remove the background noise to some extent, they leave musical residual noise, which makes speech unpleasant. Many algorithms proposed for reducing the residual noise [2] can sometimes distort the speech signal.

We inject music to improve the pleasantness of speech signal using characteristics of the musical signal. The injection is controlled by two factors. One is based on spectral information of the speech signal, the second is chosen empirically.

Pleasantness is a subjective measure, which depends on individual sensation. Although listening tests give direct evaluation, such tests are often costly and complicated. As in [2], we set up a psychoacoustic model for pleasantness considering four parameters, including loudness, sharpness,

roughness and tonality. This model makes the subjective evaluation easier and more efficient.

This paper is organized as follows. In Section 2, the general speech enhancement model is described. Our method is presented in Section 3. Calculation and meaning of different types of psychoacoustic metrics are introduced in Section 4. Section 5 provides our results.

2. PROBLEM STATEMENT

In real environment, the need to enhance speech signals arises in many situations where the speech signal is originated from a noisy location or is affected by noise over a communication channel. A simple model of speech enhancement is shown in Fig. 1. s_0 is the clean speech. s_1 is the noisy speech, which is polluted by additive noise n . After initial noise reduction stage, we obtain s_2 with lowered background noise. However, the residual noise in s_2 may be annoying. In this paper, we will inject music to s_2 in order to obtain s_3 with improved pleasantness.

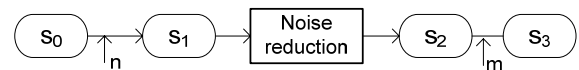


Fig.1. Model for speech enhancement

3. PROPOSED SUBJECTIVE SPEECH ENHANCEMENT METHOD

The subjective speech enhancement method we propose consists of the following two steps.

In the first step, the noisy speech is processed by a traditional noise reduction algorithm. In this paper, we adopt spectral subtraction, which is computationally simple as it only involves Fourier Transform and its inverse. It is based on a simple principle; assuming additive noise, one can obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The noise spectrum can be estimated and updated during the speech less periods. The enhanced signal

is obtained by computing inverse Discrete Fourier Transform (IDFT) of the estimated signal spectrum using the phase of noisy signal.

In the second step, we inject the music controlled by two factors in time domain. That is, we define speech signal with improved pleasantness by

$$y(n, i) = x(n, i) + c_1 \cdot c_0(i) \cdot m(n, i), \quad (1)$$

where n is the discrete-time index, i is the short-time frame index for speech signal, $x(n, i)$ is the processed signal by spectral subtraction, $m(n, i)$ is the music signal, and $y(n, i)$ is the expected signal with enhanced pleasantness, $c_0(i)$ is the Music Injection Base Factor (MIBF) which is calculated by Equation (2), c_1 is the Music Injection Modified Factor (MIMF) which is chosen empirically.

$$c_0(i) = \sqrt{\frac{\sqrt{S_x(i) \cdot S_n(i)}}{S_m(i)}}, \quad (2)$$

where $S_x(i)$, $S_m(i)$ and $S_n(i)$ are the average power spectra of the processed speech signal, music signal and the estimated noise in the i th short time frame. The time-domain process is illustrated in Fig. 2. With (2), we want to make sure the power spectrum level (dB) of the injected music is between the speech and the noise.

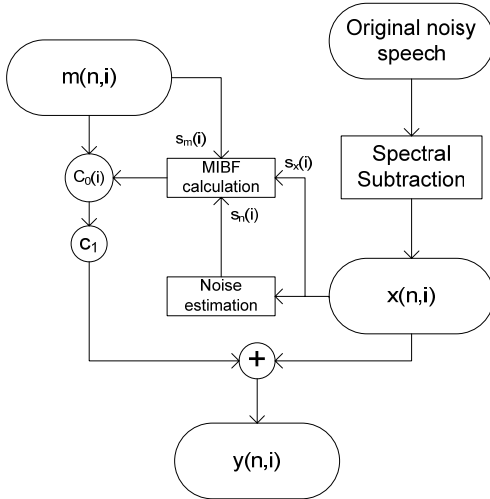


Fig.2. Flow chart of music injection process

4. PSYCHOACOUSTIC METRICS

Four psychoacoustic parameters are considered to evaluate the pleasantness in [4] [5] [6]. They are described next.

4.1. Loudness

Loudness represents the intensity sensation and is measured in units of *sones*. It can be calculated with the following function:

$$L = \int_0^{24Bark} N' dz, \quad (3)$$

where L is the loudness, N' is the “specific loudness”, i.e., the loudness in a specific critical band, which is measured in units of *sones/bark*.

4.2. Sharpness

Sharpness describes the high frequency content of a sound and is measured in *acum*. It is defined by

$$S = 0.11 \frac{\int_0^{24Bark} N' g(z) z dz}{\int_0^{24Bark} N' dz}, \quad (4)$$

Where, $g(z)$ is weighting factor defined as

$$g(z) = \begin{cases} 1; & z \leq 16 \\ 0.066 \cdot e^{0.171z}; & z > 16 \end{cases}. \quad (5)$$

4.3. Roughness

Roughness represents the human perception of temporal variations of sounds and is measured in *asper*. The expression for roughness is

$$R = 0.3 \frac{f_{mod}}{kHz} \int_0^{24Bark} \frac{\Delta L_E(z) dz}{dB / Bark}, \quad (6)$$

Where, ΔL_E is the change in the sensation level in dB, f_{mod} is the modulation frequency of the sound.

4.4. Tonality

Tonality is concerned with the tonal prominence of a sound. In [5], it is suggested that tonality has to be judged subjectively. However, Spectrum Flatness Measure (SFM) can be used to approximate tonality [3]. High SFM indicates that the spectrum has a similar amount of power in all spectral bands - this would sound similar to white noise, and the spectrum would appear relatively flat and smooth. It is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum, i.e.:

$$SFM = \frac{\sqrt[N]{\prod_{k=0}^{N-1} P(k)}}{\frac{1}{N} \sum_{k=0}^{N-1} P(k)}. \quad (7)$$

where $P(k)$ is the magnitude of k th DFT sample. Using SFM, the totality T is calculated by

$$SFM_{db} = 10 \cdot \log_{10}(SFM) \quad (8)$$

$$T = \min\left(\frac{SFM_{db}}{-60}, 1\right). \quad (9)$$

4.5 Pleasantness

In [6], an empirical model based on foregoing four factors is proposed for measuring the pleasantness quantitatively as follows.

$$P = e^{-0.55R} e^{-0.113S} (1.24 - e^{-2.2T}) e^{(-0.023L)^2}. \quad (10)$$

where L , S , R , and T are given by (3), (4), (6), and (9), respectively.

5. RESULTS

Noisy speech was obtained by adding the clean speech signal with the background noise. The clean speech is selected from IEEE corpus database [1]. The IEEE database is used because it contains phonetically-balanced sentences with relatively low word-context predictability. Because the sentence in the database is very short (about 3 seconds), we combine 3 or 4 sentences as one test sentence, which is better to evaluate the effect of music injection. With regards to the background noise, we choose babble noise from the SPIB noise database from Rice University (http://spib.rice.edu/spib/select_noise.html). SNR of the synthesized noisy speech is 5dB and the sampling frequency is 44.1kHz. The noise reduction is conducted on 20-ms duration frames of noisy signal with 50% overlap between frames. Piano music is used as an injection signal.

As expressed in (1) (2), $c_0(i)$ is calculated and updated in each short-time frame. We change the value of MIMF c_1 and calculate the corresponding loudness, sharpness, roughness and tonality. With these four factors we get the pleasantness based on (10).

We choose 20 discrete values for MIMF from 1 to 20 and conducted several simulations. Test is repeated for three different test sentences (each one is about 10 seconds) and the average results are obtained.

To verify the validity of the pleasantness model, subjective tests were conducted to evaluate the performance of music injection. 10 listeners are chosen to score the testing pairs, which include speech without music injection, 1 second silence, and speech with music injection based on the

following scoring rule: 1-Very Bad, 2-Bad, 3-Fair, 4-Good, 5-Excellent.

We choose 3 different types of instrumental music for our test: piano, bassoon and flute. The relationships between the MIMF c_1 and four psychoacoustic descriptors, the pleasantness values are shown in Fig. 3- 8. It is indicated that piano and bassoon music improve the pleasantness, however the flute music degrades it. From Fig. 4, 6 and 8, we can see the subjective test results match the objective evaluation results.

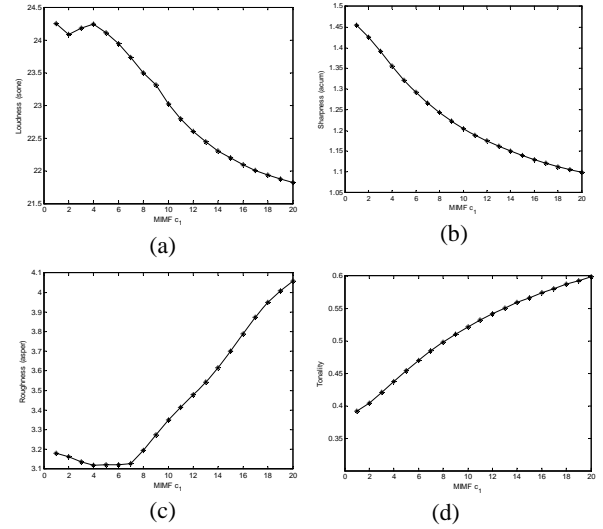


Fig. 3. Piano: Influence of MIMF on loudness (a), sharpness (b), roughness (c) and tonality (d). Open circles show subjective listening scores obtained with human listeners.

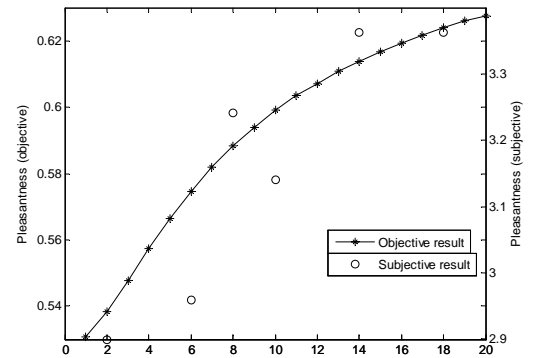


Fig. 4. Piano: Influence of MIMF on pleasantness for both objective and subjective evaluation. Open circles show subjective listening scores obtained with human listeners.

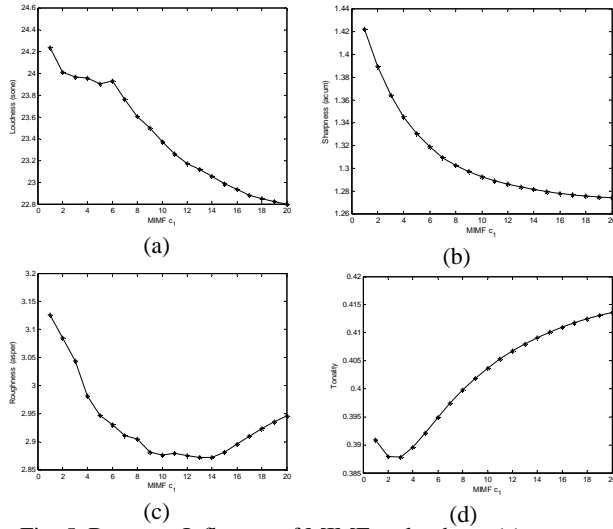


Fig. 5. Bassoon: Influence of MIMF on loudness (a), sharpness (b), roughness (c) and tonality (d).

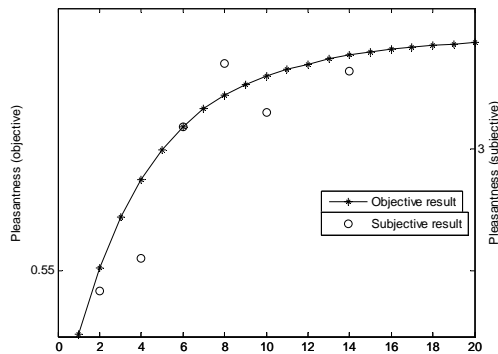


Fig. 6. Bassoon: Influence of MIMF on pleasantness for both objective and subjective evaluation. Open circles show subjective listening scores obtained with human listeners.

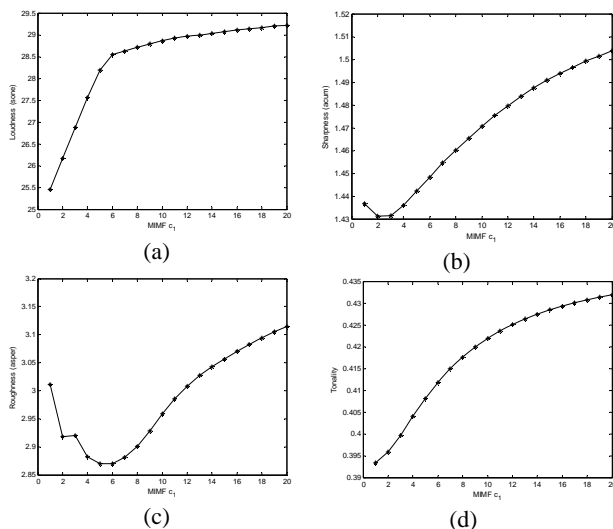


Fig. 7. Flute: Influence of MIMF on loudness (a), sharpness (b), roughness (c) and tonality (d).

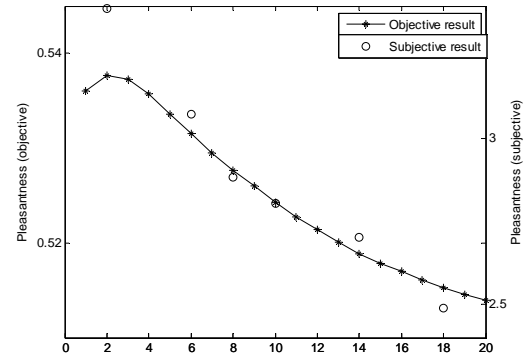


Fig. 8. Flute: Influence of MIMF on pleasantness for both objective and subjective evaluation. Open circles show subjective listening scores obtained with human listeners.

6. CONCLUSION

This paper proposes a new method from the subjective view point to make speech enhancement in terms of pleasantness. Music is injected and controlled by two factors, MIBF and MIMF. MIBF is defined by the spectral characteristics of speech and estimated noise. MIMF is chosen experimentally. The Aures' pleasantness model [6] is adopted to evaluate the performance. As simulation results show, pleasantness is improved with the injection of piano and bassoon music, although it is degraded by flute music. In the meantime, subjective tests verify the objective pleasantness model we used.

7. REFERENCES

- [1] P.C. Loizou, "Speech enhancement – Theory and Practice," Taylor&Francis, 2007.
- [2] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1998, pp. 397–400
- [3] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
- [4] H. Fastl, "The psychoacoustics of sound-quality evaluation," *Acta Acustica*, vol. 83, pp. 754–764, 1997.
- [5] E. Zwicker and H. Fastl, "Psychoacoustic – Facts and Models," Springer-Verlag, 2nd edition, 1999
- [6] W. Aures, "Berechnungsverfahren für den Wohlklang beliebiger Schallsignale, ein Beitrag zur gehörbezogenen Schallanalyse", Ph.D. thesis, Munich University, 1994.