Effect of Analysis Window Duration on Speech Intelligibility

Effect of Analysis Window Duration on Speech Intelligibility

Kuldip Paliwal, *Member, IEEE*, and Kamil Wójcicki

***Abstract*—In this letter, we investigate the effect of the analysis window duration on speech intelligibility in a systematic way. In speech processing, the short-time magnitude spectrum is believed to contain the majority of the intelligible information. Consequently, in our experiments, we construct speech stimuli based purely on the short-time magnitude spectrum. We conduct subjective listening tests in the form of a consonant recognition task to assess intelligibility as a function of analysis window duration. In our investigations, we also employ three objective speech intelligibility measures based on the speech transmission index (STI). The experimental results show that the analysis window duration of 15–35 ms is the optimum choice when speech is reconstructed from the short-time magnitude spectrum.**

***Index Terms*—Analysis window duration, magnitude spectrum, speech intelligibility, speech transmission index (STI).**

I. Introduction

ALTHOUGH speech is nonstationary, it can be assumed quasi-stationary and, therefore, can be processed through the short-time Fourier analysis. The short-time Fourier transform (STFT) of a speech signal $s(t)$ is given by

$$S(t,f)=\int_{-\infty}^{\infty} s(\tau)w(t-\tau)\mathrm{e}^{-j2\pi f\tau}d\tau \tag{1}$$

where $w(t)$ is an analysis window function of duration T_w. In speech processing, the Hamming window function is typically used and its width is normally 20–40 ms. The short-time Fourier spectrum, $S(t,f)$, is a complex quantity and can be expressed in polar form as

$$S(t,f)=|S(t,f)|\mathrm{e}^{j\psi(t,f)} \tag{2}$$

where $|S(t,f)|$ is the short-time magnitude spectrum and $\psi(t,f)=\angle S(t,f)$ is the short-time phase spectrum. The signal $s(t)$ is completely characterized by its magnitude and phase spectra.[1]

The rationale for making the window duration 20–40 ms comes from the following qualitative arguments. When making the quasi-stationarity assumption, we want the speech analysis segment to be stationary. As a result, we cannot make the speech analysis window too large; otherwise, the signal within the window will become nonstationary. From this consideration, the window duration should be as small as possible. However, making the window duration small also has its disadvantages. One disadvantage is that if we make the analysis duration smaller, then the frame shift decreases and thus the frame rate increases. This means we will be processing a lot more information than necessary, thus increasing the computational complexity. The second disadvantage of making the window duration small is that the spectral estimates will tend to become less reliable due to the stochastic nature of the speech signal. The third reason why we cannot make the analysis window too small is that in speech processing, the typical range of pitch frequency is between 80 and 500 Hz. This means that a typical pitch pulse occurs every 2 to 12 ms. If the duration of the analysis window is smaller than the pitch period, then the pitch pulse will sometimes be present and at other times absent. When the speech signal is voiced in nature, the location of pitch pulses will change from frame to frame under pitch-asynchronous analysis. To make this analysis independent of the location of pitch pulses within the analysis segment, we need a segment length of at least two to three times the pitch period. The above arguments are normally used to justify the analysis window duration of around 20–40 ms. However, they are all qualitative arguments and do not tell us exactly what the analysis segment duration should be.

In this letter, we propose to investigate a systematic way of arriving at an optimal duration of an analysis window. We want to do so in the context of typical speech processing applications. The majority of these applications utilize only the short-time magnitude spectrum information. For example, speech and speaker recognition tasks use cepstral coefficients as features which are based solely on the short-time magnitude spectrum. Similarly, typical speech enhancement algorithms modify only the magnitude spectrum and leave the noisy phase spectrum unchanged. For this reason, in our investigations, we employ the analysis-modification-synthesis (AMS) framework where, during the modification stage, only the short-time magnitude spectrum is kept, while the short-time phase spectrum is discarded by randomizing its values. In our experiments, we investigate the effect of the duration of an analysis segment used in the short-time Fourier analysis to find out what window duration gives the best speech intelligibility under this framework. For this purpose, both subjective and objective speech intelligibility measures are employed. For subjective evaluation, we conduct listening tests using human listeners in a consonant recognition task. For objective evaluation, we employ three speech-based derivatives of a popular objective speech intelligibility measure, namely, the speech transmission index (STI).

The remainder of this letter is organized as follows. Section II describes the AMS procedure used to construct stimuli files for

Manuscript received May 03, 2008; revised August 12, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Brian Kan-Wing Mak.

The authors are with the Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Nathan QLD 4111, Australia (e-mail: k.paliwal@griffith.edu.au; k.wojcicki@griffith.edu.au).

Digital Object Identifier 10.1109/LSP.2008.2005755

[1]In our discussions, when referring to the magnitude or phase spectra, the short-time modifier is implied unless otherwise stated.

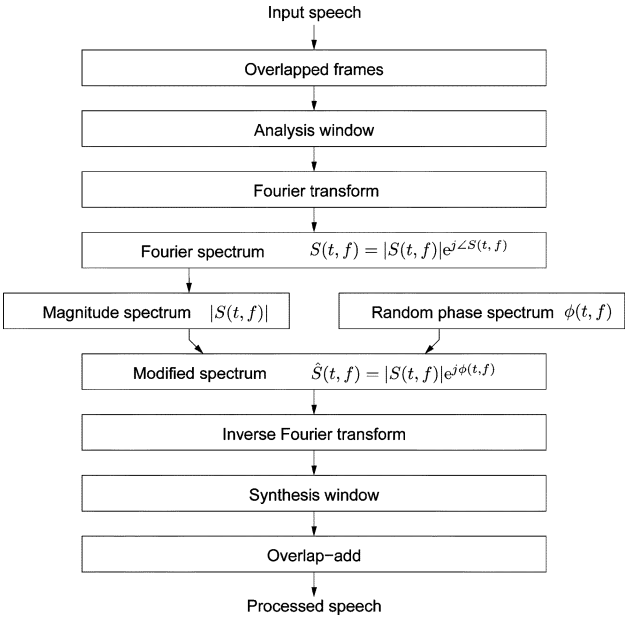

Fig. 1. Procedure used for stimulus construction.

our experiments. Section III provides details of the subjective listening tests. Section IV outlines the objective evaluation procedure. Results and discussion are presented in Section V.

II. Analysis-Modification-Synthesis

The aim of the present study is to determine the effect that the duration of an analysis segment has on speech intelligibility, using a systematic, quantitative approach. Since the majority of speech processing applications utilize only the short-time magnitude spectrum, we construct stimuli that retain only the magnitude information. For this purpose, the AMS procedure, shown in Fig. 1, is used. In the AMS framework, the speech signal is divided into overlapped frames. The frames are then windowed using an analysis window, $w(t)$, followed by the Fourier analysis, and spectral modification. The spectral modification stage is where only the magnitude information is retained. The phase spectrum information is removed by randomizing the phase spectrum values. The resulting modified STFT is given by

$$\hat{S}(t,f) = |S(t,f)|e^{j\phi(t,f)} \tag{3}$$

where $\phi(t,f)$ is a random variable uniformly distributed between 0 and 2π. Note that when constructing the random phase spectrum, the antisymmetry property of phase spectrum should be preserved. The stimulus, $\hat{s}(t)$, is then constructed by taking the inverse STFT of $\hat{S}(t,f)$, followed by synthesis windowing and overlap-add (OLA) reconstruction [1]–[4]. We refer to the resulting stimulus as magnitude-only stimulus, since it is reconstructed by using only the short-time magnitude spectrum.[2]

III. Subjective Experiment

This section describes subjective measurement of speech intelligibility as a function of analysis window duration. For this purpose, human listening tests are conducted, in which consonant recognition performance is measured.

A. Recordings

Six stop consonants, [b, d, g, p, t, k], were selected for the human consonant recognition task. Each consonant was placed in a vowel-consonant-vowel (VCV) context within the "*Hear aCa now*" carrier sentence.[3] The recordings were carried out in a silent room using a SONY ECM-MS907 microphone. Four speakers were used: two males and two females. Six recordings per speaker were made, giving a total of 24 recordings. Each recording lasted approximately 3 s, including leading and trailing silence portions. All recordings were sampled at $F_s = 16$ kHz with 16-bit precision.

B. Stimuli

The recordings were processed using the AMS procedure detailed in Section II. The Hamming window was employed as the analysis window function. Twelve analysis window durations were investigated ($T_w = 1$, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, and 2048 ms). The frame shift was set to $T_w/8$ ms and the FFT analysis length was set to $2N$, where $N(= T_w F_s)$ is the number of samples in each frame. These settings were chosen to minimize aliasing effects. For a detailed look at how the choice of the above parameters affects subjective intelligibility, we refer the reader to [6] and [7]. The modified Hanning window [4] was used as the synthesis window. The original recordings (reconstructed without spectral modification) were also included. Overall, 13 different treatments were applied to the 24 recordings, resulting in the total of 312 stimuli files. Example spectrograms of original as well as processed stimuli are shown in Fig. 4.

C. Subjects

For listeners, we used twelve English-speaking volunteers, with normal hearing. None of the listeners participated in the recording of the stimuli.

D. Procedure

The listening tests were conducted in isolation, over a single session, in a quiet room. The task was to identify each carrier utterance as one of the six stop consonants. The listeners were presented with seven labeled options on a digital computer, with the first six corresponding to the six stop consonants and the seventh being the null response. The subjects were instructed to choose the null response only if they had *no idea* as to what the embedded consonant might have been. The stimuli audio files were played in a randomized order and presented over closed circumaural headphones (SONY MDR-V500) at a comfortable listening level. Prior to the actual test, the listeners were familiarized with the task in a short practice session. The entire sitting lasted approximately half an hour. The responses were collected via a keyboard. No feedback was given.

IV. Objective Evaluation

In this section, our aim is to investigate the effect of the analysis window duration on speech intelligibility using objective measures. For this purpose, we employ the STI as the

[2]Although we remove the information about the short-time phase spectrum by randomizing its values and keep the magnitude spectrum, the phase spectrum component in the reconstructed speech cannot be removed to a 100% perfection [5].

[3]For example, for the consonant [g], the utterance is "*Hear aga now*".

performance metric [8]. STI measures the extent to which slow temporal intensity envelope modulations are preserved in degraded listening environments [9]. It is these slow intensity variations that are important for speech intelligibility. In the present work, we employ the speech-based STI computation procedure where speech signal is used as a probe. Under this framework, the original and processed speech signals are passed separately through a bank of seven octave band filters. Each filtered signal is squared and low pass filtered (with cutoff frequency of 32 kHz) to derive the temporal intensity envelope. The power spectrum of the temporal intensity envelope is subjected to one-third octave band analysis. The components over each of the 14 one-third octave band intervals (with centers ranging from 0.63 to 12.7 Hz) are summed, producing 98 modulation indices. The resulting modulation spectrum of the original speech, along with the modulation spectrum of the processed speech, can then be used to compute the modulation transfer function (MTF), which in turn is used to compute STI. In this work, three different approaches are employed for the computation of the MTF. The first approach is by Houtgast and Steeneken [10], the second is by Drullman *et al.* [11], and the third is by Payton *et al.* [12]. The details of MTF and STI computations are given in [13]. The objective evaluation is performed on the stimuli files used in the subjective experiment (see Section III-B).

V. Results and Discussion

In the subjective experiment, described in Section III, we have measured consonant recognition performance through human listening tests. We refer to the results of these measurements as subjective intelligibility scores. The subjective intelligibility scores (along with their standard error bars) are shown in Fig. 2(a) as a function of analysis window duration. The following observations can be made based on these results. For short analysis window durations, the subjective intelligibility scores are low. The scores increase with an increase in analysis window length, but at long window durations, the subjective intelligibility scores start to decrease. It is important to note that Fig. 2(a) shows a peak for analysis window durations between 15 and 35 ms.

Section IV outlines an objective evaluation of speech intelligibility. We refer to the results of this evaluation as objective intelligibility scores. The objective intelligibility scores as a function of analysis window length are shown in Fig. 2(b). The objective results show a trend similar to that of the subjective results. Although, in the objective case, the peak is not as pronounced, it can be seen to lie between 8 and 40 ms. Note that all three speech-based STI measures display a similar trend.

Mean speech-based STI scores as a function of subjective intelligibility scores, as well as least-squares lines of best fit and correlation coefficients, are shown in Fig. 3. All three STI derivatives were found to have a statistically significant correlation with subjective intelligibility scores at a 0.0001 level of significance using correlation analysis [14]. This indicates that the three STI measures can be used to predict subjective intelligibility.

Based on subjective as well as objective intelligibility scores, it can be seen that the optimum window duration for speech analysis is around 15–35 ms. For speech applications based solely on the short-time magnitude spectrum, this window duration is expected to be the right choice. This duration has been recommended in the past on the basis of qualitative arguments.

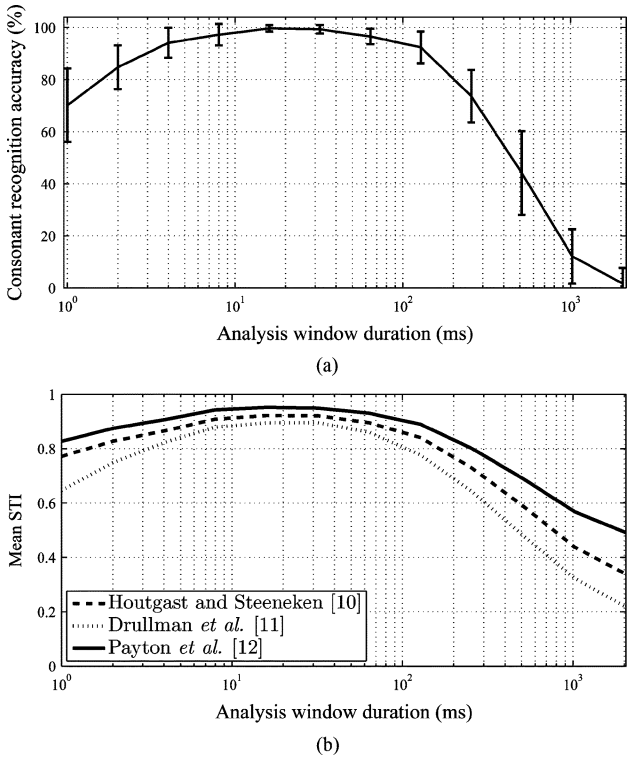

Fig. 2. Experimental results. (a) Subjective intelligibility scores in terms of consonant recognition accuracy (%). (b) Objective intelligibility scores in terms of mean speech-based STI. Objective scores are shown for the following methods: Houtgast and Steeneken method [10]—broken line, Drullman *et al.* method [11]—dotted line, and Payton *et al.* method [12]—solid line.

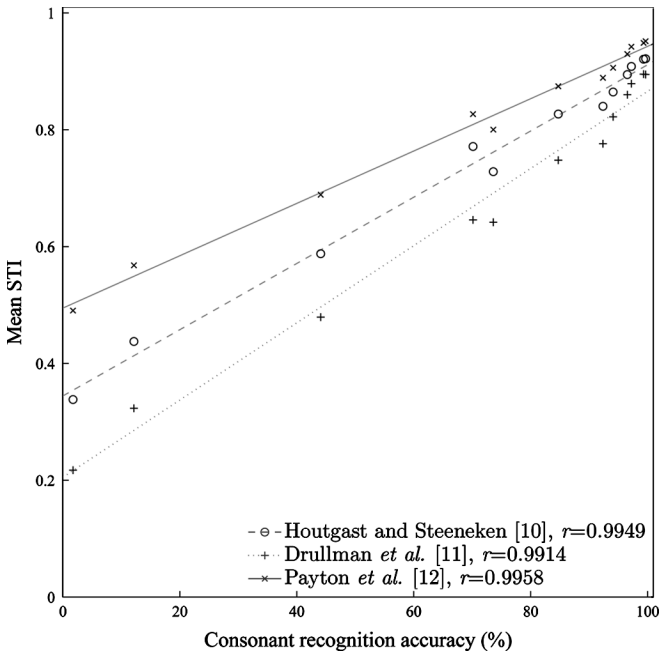

Fig. 3. Objective intelligibility scores in terms of mean speech-based STI versus subjective intelligibility scores in terms of consonant recognition accuracy (%). Correlation coefficients, r, as well as least-squares lines of best fit are also shown for each of the STI-based methods.

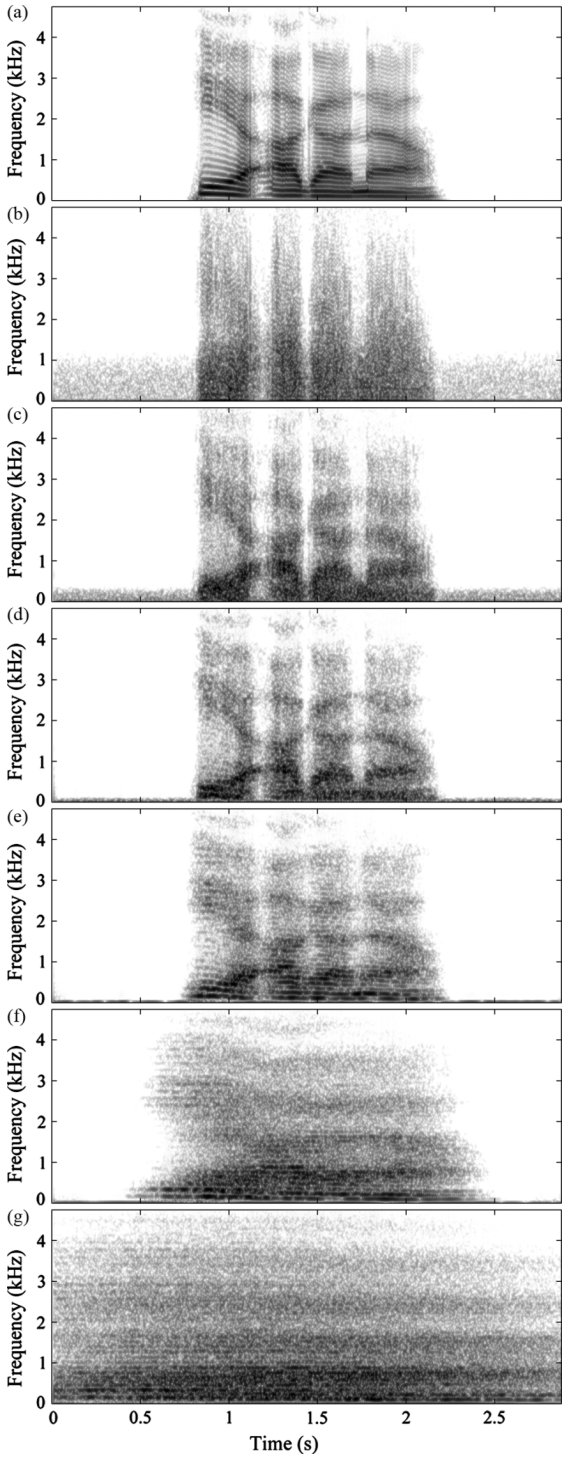

Fig. 4. Spectrograms of an utterance "*Hear aga now*", by a male speaker. (a) Original speech (passed through the AMS procedure with no spectral modification). (b–g) Processed speech—magnitude-only stimuli for different analysis window durations: (b) 2 ms, (c) 8 ms, (d) 32 ms, (e) 128 ms, (f) 512 ms, and (g) 2048 ms.

However, in the present work, the similar optimal segment length was obtained through a systematic study of subjective and objective intelligibility of speech stimuli, reconstructed using only the short-time magnitude spectrum. To the best of our knowledge, this is the first attempt to quantify the window duration on the basis of subjective intelligibility scores.

VI. Conclusion

In this letter, the effect of the analysis window duration on speech intelligibility was investigated in a systematic way. Subjective evaluation in the form of human listening tests comprising of a consonant recognition task were conducted. In addition to the subjective evaluation, three speech-based variants of the STI objective speech intelligibility measure were also employed. The experimental results show that the analysis window duration of 15–35 ms is the optimum choice when a speech signal is reconstructed from its short-time magnitude spectrum only.

References

[1] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. PROC-65, no. 11, pp. 1558–1564, Nov. 1977.

[2] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 1, pp. 99–102, Feb. 1980.

[3] M. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 364–373, Jun. 1981.

[4] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[5] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. Amer.*, vol. 110, no. 3, pp. 1628–1640, Sep. 2001.

[6] K. Paliwal and L. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Commun.*, vol. 45, no. 2, pp. 153–170, Feb. 2005.

[7] L. Alsteris and K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digit. Signal Process.*, vol. 17, pp. 578–616, May 2007.

[8] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[9] K. Payton and L. Braida, "A method to determine the speech transmission index from speech waveforms," *J. Acoust. Soc. Amer.*, vol. 106, pp. 3637–3648, Dec. 1999.

[10] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Amer.*, vol. 77, no. 3, pp. 1069–1077, Mar. 1985.

[11] R. Drullman, J. Fresten, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 2670–2680, May 1994.

[12] K. L. Payton, L. D. Braida, S. Chen, P. Rosengard, and R. Goldsworthy, "Computing the STI using speech as a probe stimulus," in *Past, Present and Future of the Speech Transmission Index*. Soesterberg, The Netherlands: TNO Human Factors, 2002, pp. 125–138.

[13] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[14] E. Kreyszig, *Advanced Engineering Mathematics*, 9th ed. New York: Wiley, 2006.