Preference for 20-40 ms window duration in speech analysis

Kuldip K. Paliwal, James G. Lyons and Kamil K. Wójcicki

Signal Processing Laboratory
Griffith University, Nathan, QLD 4111, Australia
{k.paliwal, j.lyons, k.wojcicki}@griffith.edu.au

ABSTRACT

In speech processing the short-time magnitude spectrum is believed to contain most of the information about speech intelligibility and it is normally computed using the short-time Fourier transform over 20-40 ms window duration. In this paper, we investigate the effect of the analysis window duration on speech intelligibility in a systematic way. For this purpose, both subjective and objective experiments are conducted. The subjective experiment is in a form of a consonant recognition task by human listeners, whereas the objective experiment is in a form of an automatic speech recognition (ASR) task. In our experiments various analysis window durations are investigated. For the subjective experiment we construct speech stimuli based purely on the short-time magnitude information. The results of the subjective experiment show that the analysis window duration of 15–35 ms is the optimum choice when speech is reconstructed from the short-time magnitude spectrum. Similar conclusions were made based on the results of the objective (ASR) experiment. The ASR results were found to have statistically significant correlation with the subjective intelligibility results.

Index Terms— Analysis window duration, magnitude spectrum, automatic speech recognition, speech intelligibility

1. INTRODUCTION

Although speech is non-stationary, 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

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

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

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

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 non-stationary. 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, which do not tell us exactly what the analysis segment duration should be.

In this paper 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 experiments are conducted. For the subjective evaluation we conduct listening tests using human listeners in a consonant recognition task. For the objective evaluation, we carry out an automatic speech recognition (ASR) experiment on the TIMIT speech corpus.

The remainder of this paper is organized as follows. Section 2 provides details of the subjective listening tests. Section 3 outlines the objective experiment. The results and discussion are presented in Section 4.

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

978-1-4244-7907-8/10/$26.00

2. 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.

2.1. 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 analysis-modification-synthesis (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

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

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, 2, 3, 4]. We refer to the resulting stimulus as magnitude-only stimulus, since it is reconstructed by using only the short-time magnitude spectrum.[2]

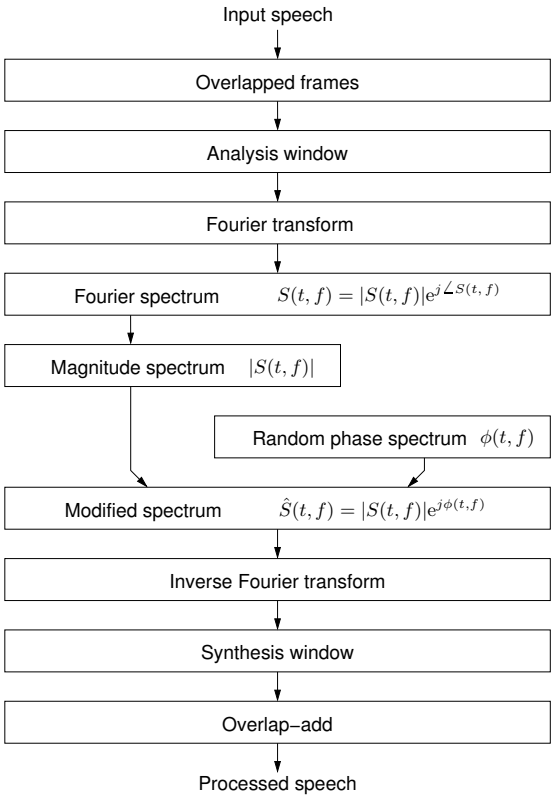

Fig. 1. Procedure used for stimulus construction.

2.2. 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 three seconds, including leading and trailing silence portions. All recordings were sampled at F_s =16 kHz with 16-bit precision.

2.3. Stimuli

The recordings were processed using the AMS procedure detailed in Section 2.1. The Hamming window was employed as the analysis window function. Ten analysis window durations were investigated (T_w = 1, 2, 4, 8, 16, 32, 64, 128, 256 and 512 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, 7]. The modified Hanning window [4] was used as the synthesis window. The original recordings (reconstructed without spectral modification) were also included. Overall, 11 different treatments were applied to the 24 recordings, resulting in the total of 264 stimuli files. Example spectrograms of original as well as processed stimuli are shown in Fig. 4.

2.4. Subjects

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

2.5. 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.

[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*".

3. OBJECTIVE EXPERIMENT

This section provides the details of our investigation of the importance of the analysis window duration in the context of a popular speech processing application, namely automatic speech recognition (ASR). For this purpose, an ASR experiment was conducted on the TIMIT speech corpus [8]. The TIMIT corpus is sampled at 16 kHz and consists of 6300 utterances spoken by 630 speakers. The corpus is separated into training and testing sets. For our experiments the *sa** utterances, which are the same across all speakers, were removed from both the training and testing sets to prevent biasing the results. The full train set, consisting of 3696 utterances from 462 speakers, was used for training, while the core test set, consisting of 192 utterances from 24 speakers, was used for testing. Both training and testing was performed on clean speech. For the our experiment we employed the hidden Markov model toolkit (HTK) [9]. A HTK-based triphone recognizer, with 3 states per HMM and 8 Gaussian mixtures per state, was used. The features used were mel-frequency cepstral coefficients [10] with energy as well as the first and second derivatives (39 coefficients total). Various analysis window durations were investigated. Cepstral mean subtraction was applied. A bigram language model was used. The training phoneme set, which consisted of 48 phonemes, was reduced to 39 for testing purposes (as in [11]). Phoneme recognition results are quoted in terms of correctness percentage [9].

4. RESULTS AND DISCUSSION

In the subjective experiment, described in Section 2, 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.

The results of the ASR experiment, detailed in Section 3, are shown in Fig. 2(b). We refer to these results as objective scores. The objective results show a trend similar to that of the subjective results. Although, in the objective case, the peak is wider and it can be seen to lie between 15 and 60 ms.

The objective scores as a function of subjective intelligibility scores, as well as least-squares line of best fit and correlation coefficient, are shown in Fig. 3. The objective scores were found to have a statistically significant correlation with subjective intelligibility scores at a 0.0001 level of significance using correlation analysis [12]. This indicates that ASR can be used to predict subjective intelligibility.

Based on subjective as well as objective results, 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. 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.

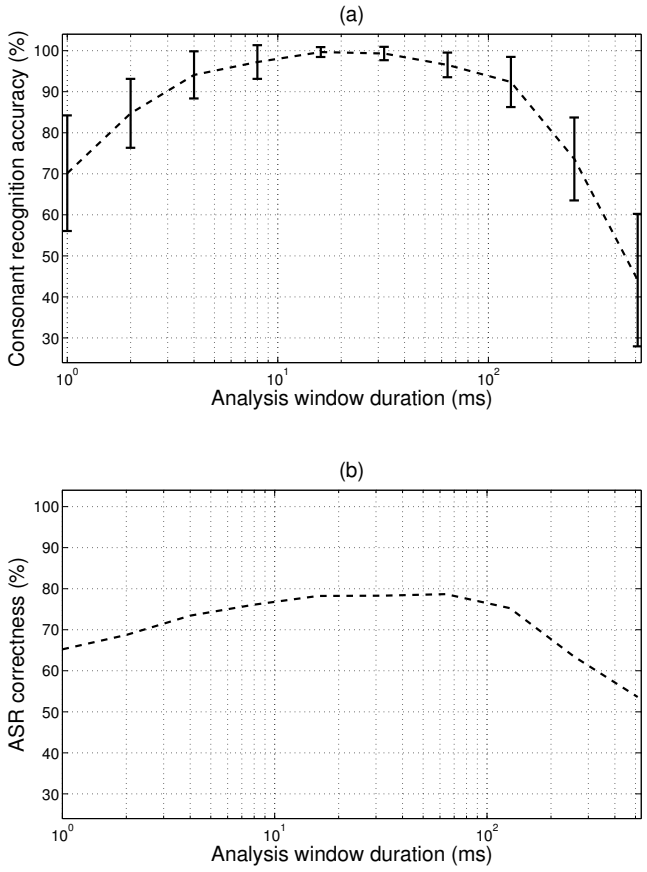

Fig. 2. Experimental results for: (a) subjective intelligibility tests in terms of consonant recognition accuracy (%); and (b) automatic speech recognition in terms of correctness (%).

5. CONCLUSION

In this paper, the effect of the analysis window duration on speech intelligibility was investigated in a systematic way. Two evaluation methods were employed, subjective and objective. The subjective evaluation was based on human listening tests that comprised of a consonant recognition task, while for the objective evaluation an ASR experiment was conducted. 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.

6. REFERENCES

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

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

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

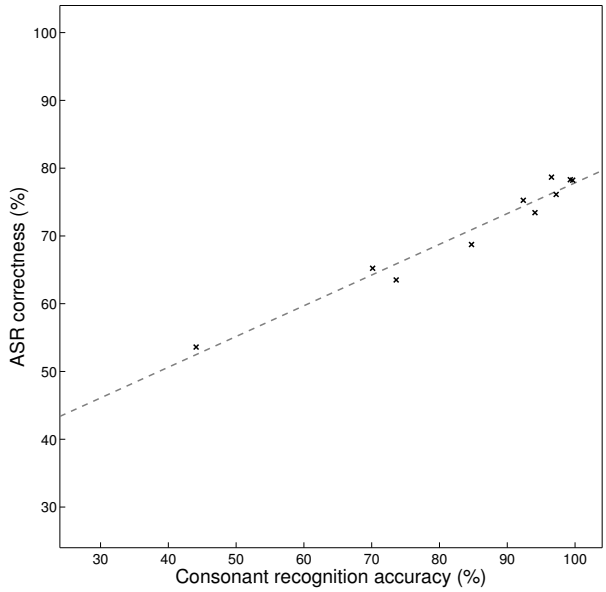

Fig. 3. Automatic speech recognition results in terms of ASR correctness (%) versus subjective intelligibility scores in terms of consonant recognition accuracy (%). Least-squares line of best fit is also shown. Correlation coefficient: $r = 0.9810$.

[4] D.W. Griffin and J.S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, 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. Am.*, vol. 110, no. 3, pp. 1628–1640, 2001.

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

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

[8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27403–+, Feb. 1993.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, Cambridge University Engineering Department, 3.4 edition, 2006.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[11] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[12] E. Kreyszig, *Advanced Engineering Mathematics*, Wiley, 9th edition, 2006.

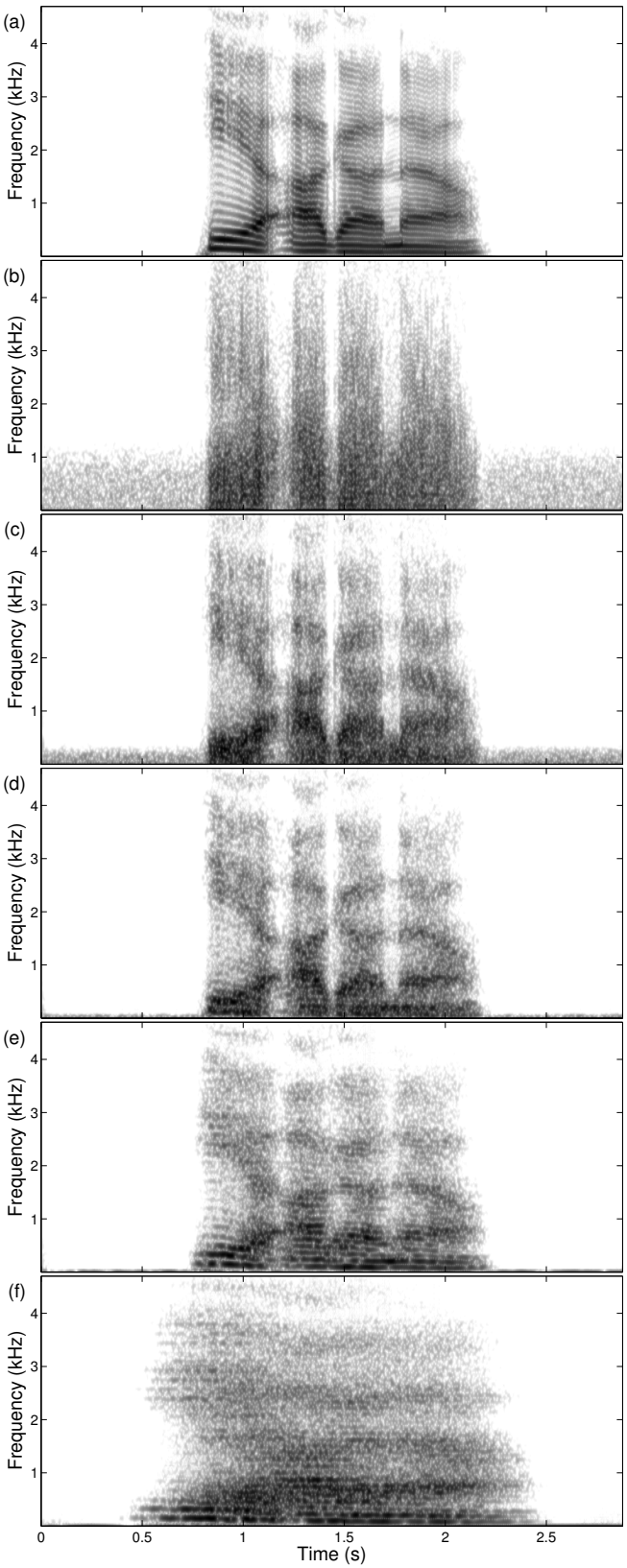

Fig. 4. Spectrograms of an utterance "*Hear aga now*", by a male speaker: (a) original speech (passed through the AMS procedure with no spectral modification); (b–f) processed speech – magnitude-only stimuli for different analysis window durations: (b) 2 ms; (c) 8 ms; (d) 32 ms; (e) 128 ms; and (f) 512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