Effective Speech Processing for Various Impaired Listeners

Takayuki Arai, Keiichi Yasu and Nao Hodoshima

Department of Electrical and Electronics Engineering
Sophia University, Japan
arai@sophia.ac.jp

Abstract

Normal hearing listeners are able to understand speech with different types of degradation, because speech has redundancy in the spectro-temporal domains. On the other hand, hearing impaired listeners have less such capability. Because of this, speech signal processing for hearing impairment needs to preserve important landmarks when enhancing a speech signal. The hearing impairments are characterized by high-frequency hearing loss, increase in the threshold of hearing, compression in the dynamic range, severity of temporal masking, and loss of spectral resolution due to the spread of masking. Thus, many studies on spectral and temporal enhancement have been proposed. In this paper we discussed effective speech processing techniques that we have developed to spectrally and temporally enhance speech signals for various types of impaired listeners.

For the spectral enhancement of speech, we have proposed two approaches: critical-band based frequency compression and formant enhancement. In critical-band based frequency compression, band-limited signals in each critical band are compressed along the frequency axis. For formant enhancement, we proposed a technique based on FFT and linear predictive coding. These techniques reduce the interference between adjacent subbands and augment the distance between the peak and valley in a speech spectrum. Also, the suppression in spectral power in the lower frequencies results in a reduction of the masking effect in higher frequencies.

For the temporal enhancement of speech, we proposed two approaches: modulation filtering and steady-state suppression. Both techniques emphasize the temporal dynamics of speech in the time domain. We previously found that the important frequency of temporal dynamics, or modulation frequency, for speech perception lie between 1-16 Hz, especially 4 Hz (Arai et al., JASA, 1999). Therefore, we proposed modulation filtering to emphasize these modulation frequencies. We further discussed steady-sate suppression (Arai et al., Acoust., Sci. & Tech., 2002) for improving speech intelligibility for hearing impaired listeners.

1. Introduction

Normal hearing (NH) listeners are able to understand speech with different types of modification. Warren et al. reported that little spectral information is required to identify component words when sentences were heard through narrow spectral slits [1]. Shannon et al. showed that the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for the recognition of speech [2]. Arai and Greenberg [3] showed that human listeners are tolerant of cross-channel spectral asynchrony. Thus, speech has redundancy in the spectral and temporal domains, and even if decimation occurs in both domains at the same time, speech sounds, such as shown in Fig. 1, are still intelligible.

Stevens [4] proposes a model of lexical access in which the acoustic speech signal is processed to yield a discrete representation of the speech stream in terms of a sequence of segments. In this model, the processing of the signal includes a step for detecting acoustic landmarks. While NH listeners may be able to reconstruct landmarks even if there is severe degradation in the speech signal, hearing impaired (HI) listeners have less such capability. However, there might be a chance that some modification on a speech signal, or speech enhancement, can improve its intelligibility if we carefully design speech processing to preserve such landmarks.

Sensorineural impairments are characterized by high-frequency hearing loss, increase in the threshold of hearing, compression in the dynamic range, severity of temporal masking, and loss of spectral resolution due to the spread of masking [5]. Because of this, there are many studies on spectral and temporal enhancements. For example, speech contrast enhancement has been reported as a signal enhancement technique [6,7]. Several studies on temporal enhancement have also been proposed. In [8], the temporal characteristics of speech were enhanced to compensate for temporal masking. This technique essentially enhances consonants without extracting them explicitly. Another technique [9] has attempted to make each speech segment stand out, claiming that not all HI listeners can gain by the adjustment of incoming speech signals to their dynamic range.

In this paper we discuss effective speech processing techniques that we have developed to spectrally and temporally enhance speech signals for various types of impaired listeners.

2. Spectral enhancement

Frequency selectivity is reduced in sensorineural HI patients because of damage to the cochlea [10]. Glasberg and Moore [11] measured an auditory filter of HI and NH people with a notched-noise masker and reported that HI patients have a wider auditory filter than NH. This smoothes the internal spectral representation of speech and causes poor performance of speech recognition by HI listeners.

Spectral enhancement might compensate for HI listeners' reduced spectral resolution. Several types of hearing aids exist today, which deal with this deficiency in varying degrees. Spectral contrast enhancement (e.g., [6,7]) is a method that enhances the major spectral prominences without enhancing fine-grain spectral features.

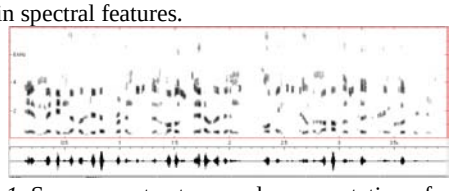

Figure 1: Sparse spectro-temporal representation of speech.

In [5], a speech signal was split into 18 critical bands, and a set of odd-numbered bands was presented to the subject's right ear, while the rest was presented to the left ear. The speech signals became clearer for both NH and HI subjects. This approach, however, is only useful when both ears have similar auditory characteristics.

Another, frequency compression, compresses the spectrum to a lower frequency to access low-frequency residual hearing [12]. Lowering the frequency of the speech in such study was done to move the speech into a frequency range where the HI patients (e.g., gradually-sloping hearing loss or steeply-sloping hearing loss) had some residual hearing. But as more severe compressions are performed, the intelligibility of the signal will tend to decrease.

We developed several new methods to enhance spectral information as follows.

2.1. Critical-band based frequency compression

Yasu et al. [13-15] proposed "critical-band based frequency compression," in which band-limited signals in each critical band were compressed along the frequency axis in light of the shape of the auditory filters of HI patients. This method maintains the peak but reduces non-peak energy within each critical band. Their method does not compress the broadband spectrum. Thus the spectral shape does not change very much and users do not need time to adapt to the sound.

In this technique, a speech signal was compressed toward the center of each critical band along the frequency axis. First, an input speech signal was divided into frames by a Hamming window. Next, the signal for each frame was transformed from the time domain to the frequency domain by FFT. After the amplitude and phase spectra of the FFT were calculated, frequency compression was done within each band. The frequency compression was done for the amplitude spectrum toward the center of each critical band along the frequency axis. The compression rate ranged from 10% to 90%. Next, the amplitude spectrum, after piece-wise frequency compression, was multiplied by the original phase spectrum. Finally, the overlap add (OLA) technique was applied to the inverse FFT (IFFT) of the product from the previous step to obtain the final signal.

Yasu et al. [16] has reported the effectiveness of critical-band based frequency compression with a profound HI listener as a case study. In this study, fifty syllables (V and CV syllables) uttered by a male speaker of Japanese were used. Because the compression rate of 50% yielded the best performance for all subjects in previous studies [13-15], the same compression rate was used. A set of stimuli consisted of 100 syllables: 50 unprocessed (control) and 50 processed syllables. After 20 sessions, a significant improvement was observed for the correct rate of syllable identification: 38.3% for the processed set, as opposed to 35.4% for the control set ($p < .05$).

2.2. Formant enhancement

Formants play an important role in speech perception, and this is crucial for HI listeners. Baer et al. [6] enhanced major spectral prominences without enhancing fine-gain spectral features that would not be resolved by a normal ear. They reported that the scores were improved by spectral enhancement, and they were improved still more by enhancement combined with amplitude compression.

Here, we proposed a formant enhancement technique based on FFT and linear predictive coding (LPC). First, an input speech signal is divided into frames by a Hamming window with the window length of 32 ms and a frame shift of 8 ms. We then perform *p*-th order LPC analysis for each frame. The sharp spectral peaks are extracted by comparing the bandwidth of a peak with a certain threshold. Finally, the extracted peaks (and their vicinities) of the FFT-based amplitude spectrum for each frame are unchanged but the rest of the spectrum is suppressed. The OLA technique is applied to the IFFT of the product of the modified amplitude and the original phase spectra.

An HI listener participated in a subjective evaluation test with a five-point scale, and the result showed that the subject scored some processed speech samples higher than the original (5.0 for such processed samples, as opposed to 4.5 for the original). This technique reduces the interference between adjacent subbands and augments the distance between peak and valley in a speech spectrum. At the same time, the suppression in spectral power in lower frequencies results in a reduction of the masking effect in higher frequencies.

3. Temporal enhancement

We have developed several techniques for temporal enhancement, including modulation filtering and steady-state suppression. They were originally designed for so-called "pre-processing," which we discuss in Sections 3.1 and 3.2. These techniques can also be applied in the speech processing of hearing aids. We discuss the effect of steady-state suppression for hearing aids in Section 3.3

3.1. Two types of processing techniques

When listening to a lecture in a large auditorium, it is often difficult to understand the speech. Comprehension may be impaired by reverberation, which is sound reflecting from the wall, interfering with direct sound. Reverberation is problematic especially for elderly people and HI listeners [17]. To prevent intelligibility degradation, we developed two types of pre-processing techniques, where signal processing is done before a signal is radiated through a loudspeaker of the public address system.

3.1.1. Modulation filtering

For pre-processing, Langhans and Strube [18] tried "modulation filtering," where they filtered temporal envelopes of subbands of the speech signal. In their study, sentence recognition tests did not show significant improvements over the original speech. We extended their attempt of modulation filtering as pre-processing [19-21].

The modulation spectrum is defined as the spectral analysis of temporal trajectories of the spectral envelopes of speech. In the clean condition, the dominant component of the modulation spectrum of continuous speech lies between 1 Hz and 16 Hz (especially 2 Hz to 8 Hz) with its peak around 4 Hz in modulation frequency. This reflects the syllabic temporal structure of speech [22,23]. The peak of the modulation spectrum shifts from around 4 Hz in clean speech to a lower modulation frequency in reverberant speech, and the modulation index decreases. This is based on the fact that the modulation transfer function (MTF) of reverberation has lowpass characteristics [24]. Performing modulation filtering

as a pre-processing technique serves to emphasize the modulation characteristics of speech prior to the influence of reverberation.

3.1.2. Steady-state suppression

In modulation filtering, linear filters are usually applied to the temporal envelope of speech signals. However, non-linear processing, such as steady-state suppression [25], may also be applied. This steady-state suppression is especially aimed to reduce the amount of "overlap-masking" of reverberation.

Overlap-masking is the main reason reverberation degrades speech intelligibility [26-28]. Because of overlap-masking, the reverberant components of one segment mask the segments that follow. As a result, the segments following the reverberating segments are harder to hear. This is especially true when the reverberating segment has more power, such as a vowel, and the subsequent has less, such as a consonant [25].

To reduce overlap-masking, Arai et al. [25] proposed "steady-state suppression." In this technique, first, an original signal is split into 1/3-octave bands. In each band the envelope is extracted. After down-sampling, the regression coefficients are calculated from the five adjacent values of the time trajectory of the logarithmic envelope of a subband. Then the mean square of the regression coefficients, *D*, is calculated. This parameter *D* is similar to what Furui proposed to measure the spectral transition [29]. After up-sampling, we define a speech portion as steady-state when *D* is less than a certain threshold. Once a portion is considered steady-state, the amplitude of the portion is multiplied by a factor less than 1.0 (this factor was set to 0.4 in the following experiments).

We compared the modulation spectra of signals with and without steady-state suppression. The modulation spectrum is derived from a frequency analysis of the temporal envelope of a band-passed signal. First, speech signals are divided into four bands and the modulation spectrum is calculated in each band. Then these modulation spectra are averaged over 24 sentences used in the perceptual experiment. When we compared the two modulation spectra, we observed that their modulation indices around 4Hz and above 10Hz are intensified by steady-state suppression in all bands (more details, see [30]).

3.2. Effect of temporal enhancement as pre-processing of speech signals

We conducted a series of experiments to investigate the effects of these techniques as pre-processing for NH and HI subjects.

3.2.1. Modulation filtering

In [19], four HI subjects with sensorineural hearing loss participated in a preference test. They rated the processed speech easier to hear than the unprocessed speech. This experiment indicates that enhancing modulation frequencies between 2 and 8 Hz improves intelligibility in reverberant environments.

In [20], the modulation filter was designed by the inverse MTF. Kitamura et al. conducted perceptual experiments with one HI subject and two subjects with NH. In [21], 31 NH subjects and one profoundly HI subject participated in the experiment with the same modulation filter as was used in [19]. These results also indicate that their method improves the intelligibility of reverberant speech.

3.2.2. Steady-state suppression

In this section, three experimental results by Hodoshima et al. [31-33] are discussed. In all experiments, speech intelligibility of two sets of speech samples with and without steady-state suppression are compared. The original speech samples consisted of nonsense consonant-vowel syllables (24 CV syllables) embedded in a Japanese carrier phrase. To achieve various reverberation conditions (reverberation time, RT, varied from 0.4 up to 1.3 s) all of the speech samples were convolved with various impulse responses before being presented to subjects.

Twenty-two NH subjects (native speakers of Japanese) participated in each of the first two experiments [31, 32]. For the comparison with and without processing, a significant improvement was obtained with the processing condition when RT was between 0.7-1.2 s.

To investigate the effect of steady-state suppression for elderly people with presbycusis, Hodoshima et al. [33] conducted a similar experiment for a male subject, age 65. In this study, a subset of the stimuli were used: RT = 0.0, 0.7, 1.0, and 1.2 s. The subject participated in the perceptual test 20 times. For the comparison of mean percent correct with and without processing, a significant improvement was obtained with processing conditions when RT is 0.7 s (unprocessed: 45.4%, processed: 50.0%) and 1.0 s (unprocessed: 41.9%, processed 45.6%). These results show that steady-state suppression prevents degrading speech intelligibility under certain reverberant conditions. These also indicate that the range of reverberation conditions in which clear improvement is observed is different in the hearing level among subjects.

3.3. Steady-state suppression as speech processing for hearing aids

Confusion in speech perception often concerns consonants, and one of the reasons might be temporal masking. Many studies apply dynamic amplitude compression, such as the automatic gain control (AGC) [34], to compensate for the recruitment phenomenon, in which the negative effect of amplitude compression has been reported [35]. In [36], modulation filtering [19-21] is applied to reduce this negative effect.

In this section, we discuss steady-state suppression as a speech processing technique for the temporal enhancement of hearing aids. As Section 3.2.2, speech intelligibility of two sets of speech samples with and without the steady-state suppression was compared. The original speech sample was obtained from the ASJ Continuous Speech Corpus (the Japanese Newspaper Article Sentences). We used eight male speakers from the corpus. Two-hundred stimuli were prepared in total (100 with processing and another 100 without processing). The same subject in [16] participated in the experiment. Performance was compared in terms of the percentage of correct morae within each sentence.

Unfortunately, the result did not show any significant difference in the performance with and without processing (55.7% without and 48.6% with the processing). This might

be due to the relatively stronger suppression, the rate of which was originally set for pre-processing; further investigations are needed.

4. Summary

We discussed several methods for spectral and temporal enhancements, including our own. There are many possible ways to combine these processing methods, yet because of their non-linearity we need to consider carefully how they may best be combined for any given task.

5. Acknowledgements

I would like to thank all of the people who helped me in various ways, especially Yôiti Suzuki, Hideki, Tachibana, Kanako Ueno, Sakae Yokoyama, Kei Kobayashi, Akiko Kusumoto, Tomoko Kitamura, Keisuke Kinoshita, Koshi Shinohara, Tsuyoshi Inoue, Masato Hishitani, Takahito Goto, Miyuki Yasuda and Noriko Ohata.

6. References

[1] Warren, R. M. et al., "Spectral redundancy: Intelligibility of sentences hear through narrow spectral slits," *Perception & Psychophysics*, 57(2):175-182, 1995.

[2] Shannon, R. V. et al., "Speech recognition with primarily temporal cues," *Science*, 270:303-304, 1995.

[3] Arai, T. and Greenberg, S., "Speech intelligibility in the presence of cross-channel spectral asynchrony," *IEEE ICASSP*, 933-936, 1998.

[4] Stevens, K. N., "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, 111(4):1872-1891, 2002.

[5] Chaudhari, D. S. and Pandey, P.C., "Dichotic presentation of speech signal with critical band filtering for improving speech perception," *IEEE ICASSP*, 6:3601-3604, 1998.

[6] Baer, T. et al., "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment," *J. Rehabil. Res. Dev.*, 30(1):49-72, 1993.

[7] Bunnell, H. T., "On enhancement of spectral contrast in speech for hearing-impaired listeners," *J. Acoust. Soc. Am.*, 88(6):2546-2556, 1990.

[8] Suzuki, R. et al., *Audiology Japan*, 34(5):335-336, 1991.

[9] Miura, M. et al., *Audiology Japan*, 38(5):469-470, 1995.

[10] Zwicker, E. and Schorn, K., "Psychoacoustical tuning curves in audiology", *Auditory*, 17:120-140, 1978.

[11] Glasberg, B. R. and Moore, B. C. J., "Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments," *J. Acoust. Soc. Am.*, 79:1020-1033, 1986.

[12] Turner, C. W. and Hurtig, R. R., "Proportional frequency compression of speech for listeners with sensorineural hearing loss," *J. Acoust. Soc. Am.*, 106:877-886, 1999.

[13] Hishitani, M. et al., "Compressing critical bands for digital hearing aids," *IHCON*, 64-65, 2002.

[14] Yasu, K. et al., "Critical-band compression method for digital hearing aids," *Forum Acusticum Sevilla*, 2002.

[15] Yasu, K. et al. "Frequency compression of critical band for digital hearing aids," *China-Japan Joint Conf. on Acoust.*, 159-162, 2002 (also, *Acoust. Sci. & Tech.*, 2004).

[16] Yasu, K. et al., "An evaluation of the critical-band compression algorithm for the wider auditory filter of hearing impaired people," *Autumn Meet. Acoust. Soc. Jpn.*, 415-416, 2003.

[17] Nabelek, A. K. and Mason, D., "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *J. Speech and Hearing*, 24:375-383, 1981.

[18] Langhans, T. and Strube, H. W., "Speech enhancement by nonlinear multiband envelope filtering," *IEEE ICASSP*, 156-159,1982.

[19] Kusumoto, A. et al., "Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired," *IEEE ICASSP*, 933-936, 2000.

[20] Kitamura, T. et al., "Designing modulation filters for improving speech intelligibility in reverberant environments," *ICSLP*, 3:586-589, 2000.

[21] Hodoshima, N. et al., "Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments," *Forum Acusticum Sevilla*, 2002.

[22] Duquesnoy, A. J. and Plomp, R., "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," *J. Acoust. Soc. Am.*, 68(2):537-544, 1980.

[23] Arai, T. et al., "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, 105(5):2783-2791, 1999.

[24] Houtgast, T. and Steeneken, H. J. M., "A review of MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, 77(6):1069-1077, 1985.

[25] Arai, T. et al., "Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments," *Acoust. Sci. & Tech.*, 23(4):229-232, 2002.

[26] Knudsen, V. O., "The hearing of speech in auditoriums," *J. Acoust. Soc. Am.*, 1:56-82, 1929.

[27] Bolt, R. H. and MacDonald, A. D., "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, 21:577-580, 1949.

[28] Nabelek, A. K. et al., "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, 86:1259-1265, 1989.

[29] Furui, S., "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, 80(4):1016-1025, 1986.

[30] Hodoshima, N. et al., "Suppressing steady-state portions of speech for improving intelligibility as pre-processing", *Technical Report of IEICE*, SP2002-65:47-51, 2002.

[31] Hodoshima, N. et al., "Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments," *China-Japan Joint Conf. on Acoust.*, 199-202, 2002 (also, *Acoust. Sci. & Tech.*, 2004).

[32] Hodoshima, N. et al., "Improving speech intelligibility by steady-state suppression as pre-processing in small to medium sized halls," *Eurospeech*, 1365-1368, 2003.

[33] Hodoshima, N. et al., 2004 (to be published).

[34] Moore, B. C. J. and Glasberg, B. R., "A compression of four methods implementing automatic gain control (AGC) in hearing aids", *Brit. J. Audiol.*, 22:93-104, 1988.

[35] Plomp, R., "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," *J. Acoust. Soc. Am.*, 83:2322-2327, 1988.

[36] Kurosawa, T. et al., "A study on validity of envelope emphasis for amplitude compression hearing aids," *Spring Meet. Acoust. Soc. Jpn.*, 459-460, 2003.