Study on important role of temporal amplitude-modulation feature of speech for auditory perception

Masashi Unoki
School of Information Science, Japan Advanced Institute of Science and Technology

Abstract: ***Amplitude modulation is a significant concept of speech transmission from speaker to listener. This concept has been separately used in the research fields of speech science, hearing, and room acoustics, but has not yet used in all consistently. This concept may be able to be used to reveal significant mechanism of human auditory perception. Recently, techniques of speech analysis and synthesis, speech enhancement, speech recognition systems, and hearing aids have been developed by various approaches such as mathematical and statistical studies. However, there is still a big gap between human auditory perception and computational models that these techniques provide due to many artifacts. To reduce this gap and to improve compatibility between human beings and computer models, significant concept should be comprehensively used in all the research fields of speech science, auditory perception, and room acoustics. This paper introduces the scheme of speech signal processing based on the concept of modulation transfer function as one of possible solutions.***

Keywords: ***Auditory filterbank, modulation filterbank, temporal amplitude envelope, noise-vocoded speech, speech intelligibility, speaker individuality***

1. Introduction

Speech is, at any time, used as one of the important and natural ways in communications for human beings to express linguistic and nonlinguistic information. In particular, nonlinguistic information such as emotion, gender, age, and speaker individuality is used for rich speech communications by human beings. Important features related to linguistic and nonlinguistic information are redundantly contained in the speech signals. Therefore, human beings can easily and correctly recognize these information even if some of the features are smeared due to noise and reverberation.

It has been studied that important acoustical features in both time and frequency domains based on the source-filter model such as fine-structure (harmonicity and periodicity), formants, spectral tilt, and temporal power fluctuation are essential for speech perception. However, recent psychoacoustical studies based on noise-vocoded speech scheme have revealed that temporal amplitude information plays an important role in speech perception [1, 2]. Shannon *et al*. reported that the presentation of modulation information of only a few acoustic bands such as noise-vocoded speech is sufficient for speech recognition [1]. In addition, Drullman *et al*. reported that cue of the temporal amplitude envelope is more important for speech perception than that of the temporal fine structure [2]. It is also well-known that low modulation frequencies, particularly below 16 Hz, play a crucial role for speech intelligibility [2].

On the one hand, amplitude modulation domain is a very important dimension in hearing, since there is modulation frequency selectivity in the auditory system [3, 4]. Modern psychoacoustical studies of temporal amplitude-modulation processing suggest that temporal amplitude envelope is processed by a modulation filterbank. Our interested question is what role of the modulation frequencies plays in hearing for temporally smeared speech signal in real environments.

One the other hand, the quality of speech transmission must be evaluated to design the required room acoustics, although many subjective experiments should be carried out to evaluate it and the costs involved are very expensive. Therefore, as a meaning of auditory perception, objective indices and measurements in room acoustics are needed to inexpensively assess the quality and intelligibility of speech. Concept of modulation transfer function (MTF) is introduced as a measure in room acoustics for assessing the effect of the enclosure on speech intelligibility [5].

In this paper, temporal amplitude-modulation features of speech for auditory perception as an important role of speech transmission in our day-life environments is introduced.

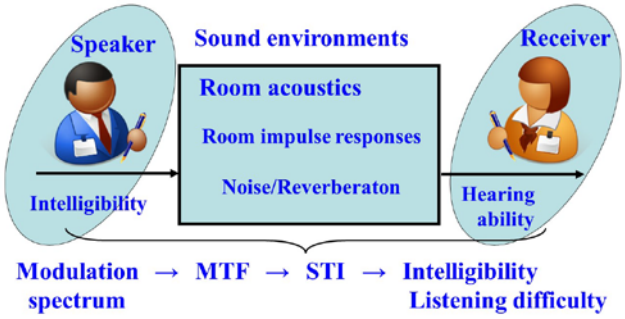

Figure 1. Block diagram for speech transmission based on concept of amplitude modulation: modulation spectrum, modulation transfer function, and modulation filterbank.

2. Concept of amplitude modulation transfer

Figure 1 shows block diagram of speech transmission base on the concept of amplitude modulation transfer: modulation spectrum of speech, modulation transfer function in room acoustics, and modulation perception on auditory filterbank.

2.1 Modulation spectrum of speech

Recent studies by Greenburg, Atlas, and Hermansky revealed that temporal amplitude envelope or its modulation spectrum conveys linguistic information of speech, as shown in Fig. 2 [6]. In particular, low frequency modulations (corresponding to 1/syllable duration) of sound have been shown to be the fundamental carriers of information in speech. Drullman *et al*. [2], for example, investigated the importance of modulation frequencies for intelligibility by applying low-pass and high-pass filters to the temporal envelopes of acoustic frequency sub-bands. They showed frequencies between 4 and 16 Hz to be important for intelligibility, with the region around 45 Hz being the most significant. In a similar study, Arai *et al*. [7] showed that applying band-pass filters between 1 and 16 Hz does not impair speech intelligibility.

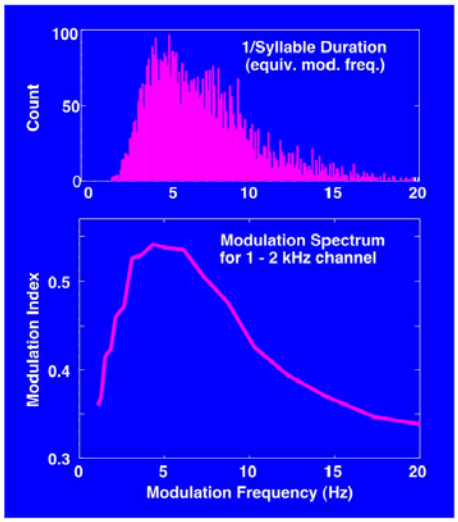

Figure 2. Modulation spectrum of speech

While the envelope of the acoustic magnitude spectrum represents the shape of the vocal tract, the modulation spectrum represents how the whole power spectrum changes as a function of time, as shown in Fig. 3 [6]. It is these temporal changes that convey most of the linguistic information (or intelligibility) of speech. In the above intelligibility studies, the lower limit of 1 Hz stems from the fact that the slow vocal tract changes do not convey much linguistic information. In addition, the lower limit helps to make speech communication more robust, since the majority of noises occurring in nature vary slowly as a function of time and hence their modulation spectrum is dominated by modulation frequencies below 1 Hz. The upper limit of 16 Hz is due to the physiological limitation on how fast the vocal tract is able to change with time.

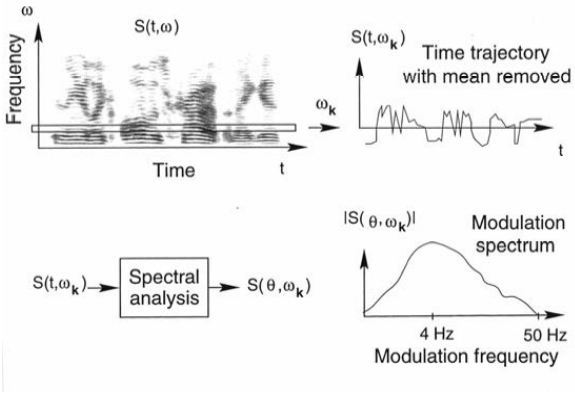

Figure 3. General scheme for deriving modulation spectrum

2.2 Modulation transfer function

Houtgast and Steeneken have proposed a method of prediction that can assess the effects of the enclosure on the intelligibility of speech in both noisy and reverberant environments by using the modulation transfer function (MTF) [5]. This concept has been used to account for the relation between the degree of modulation of the envelopes of input and output signals and the characteristics of the enclosure as shown in Fig. 4 and a way to predict the speech transfer index (STI) as shown in Fig. 4, which is strongly related to intelligibility. This concept enables noise and reverberation to be simultaneously suppressed so that there are some applications of speech signal processing that should be resolved. In this paper, a few topics of speech signal processing based on the MTF concept that the author proposed have been introduced.

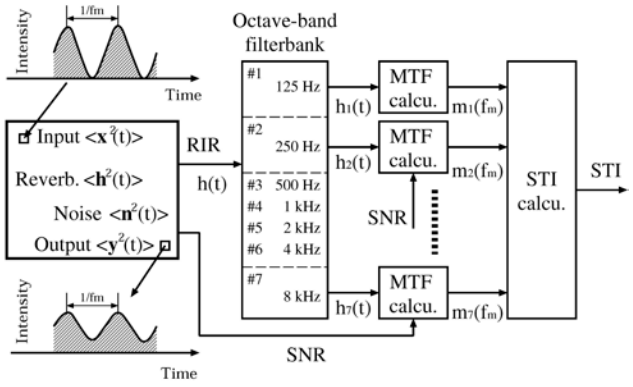

Figure 4. Block diagram of MTF and STI calculations

In this concept, noisy reverberant environments are assumed that noise property is stationary (white Gaussian noise) and reverberation property is the diffused sound field (Schroeder model). In this case, the MTF in noisy reverberant environments depends on the modulation frequency. This means the low-pass characteristics resulting from reverberation as a function of the reverberation time and the constant attenuation resulting from noise as a function of the signal to noise ratio (SNR). Most interesting point is decomposable the MTF in reverberant environments and the MTF in noisy environments separately.

2.3 Modulation filterbank

The main function of the human auditory system is to decompose sound signals into frequency components (i.e., frequency selectivity), as shown in Fig. 5. It is well known that this frequency selectivity involves nonlinear signal processing. We have been correcting the masking data of various masking situations to find nonlinear frequency selectivity [8]. Resent nonlinear auditory filterbank whose function is equivalent that of the human hearing system has been proposed [3]. Further, the same processes as half-wave rectification (HWR) and low-pass filtering (LPF) are performed as the mechanisms for inner hair cells and neural firing. In other words, the human auditory peripheral system has processes such as band division and temporal envelope extraction, as shown in Fig. 6. Amplitude modulation is a very important dimension in hearing, since there is modulation frequency selectivity in the auditory system.

There is growing psychoacoustic and physiological evidence to support the significance of the modulation domain in the analysis of speech signals. Experiments of Bacon and Grantham, for example, showed that there are channels in the auditory system which are tuned for the detection of modulation frequencies. Sheft and Yost showed that our perception of temporal dynamics corresponds to our perceptual filtering into modulation frequency channels and that faithful representation of these modulations is critical to our perception of speech. Experiments of Schreiner and Urbas showed that a neural representation of amplitude modulation is preserved through all levels of the mammalian auditory system, including the highest level of audition, the auditory cortex. Neurons in the auditory cortex are thought to decompose the acoustic spectrum into spectro-temporal modulation content, and are best driven by sounds that combine both spectral and temporal modulations.

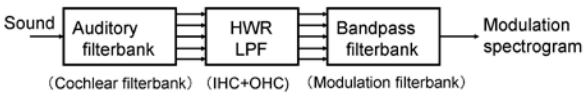

Figure 5. Derivation of modulation spectrogram

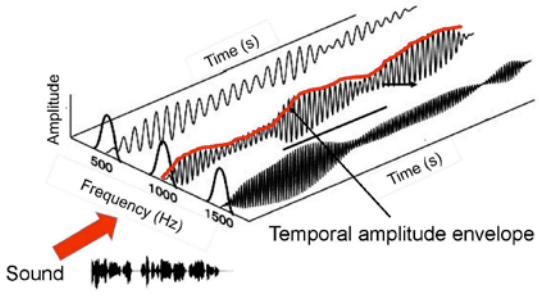

Figure 6. Temporal amplitude envelopes of sub-band signals.

3. Noise-vocoded speech scheme

Modern psychophysical models of temporal modulation processing suggest that temporal envelope is processed by a modulation filterbank [3]. Therefore, in the auditory system, modulation frequency analysis should be used to extract the linguistic and nonlinguistic information included in the temporal envelope of speech. The aim of this study is to investigate the role of auditory modulation filtering for recognizing speech intelligibility and the others such as speaker individualities and emotion in noise-vocoded speech, since natural spoken-language processing includes both speech and speaker recognition.

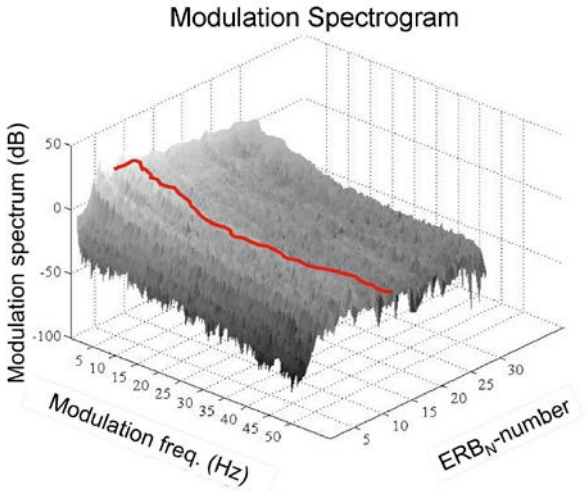

Figure 7. Example of modulation spectrogram for speech

In noise-vocoder scheme, speech sound can be resynthesized in which the temporal fine structure (TFS) of speech in sub-bands are replaced by bandlimited noise carriers while the temporal amplitude envelopes of speech are preserved. Therefore, this scheme can reveal what the role of temporal amplitude modulation for speech perception is [1].

In our experiments, speech signal was decomposed into ERB_N-bands by an auditory-motivated filterbank and temporal envelopes of ERB_N-bands were modulated with noise-carriers of the same bands. A modulation filterbank was also used to control the envelopes of octave-bands from 2 to 64 Hz [3]. We then investigated the commonalities and individual differences of modulation aspects, by analyzing modulation spectra in the all bands as shown in Fig. 7, for all stimuli from different speech contents and speakers.

For speech recognition, the results showed that speech recognition rate is drastically reduced as the maximum modulation frequency of temporal envelope lowers bellow 5 Hz [9]. In addition, for speaker recognition, the results showed that the largest variances at the modulation frequencies, lower than 15 Hz, were observed in the all bands ranges from 20 to 29 ERB_N-numbers [10]. Then, in the psychoacoustic experiments, the results showed that the speaker recognition rate is drastically reduced as the modulation frequency lowers below 8 Hz. These results suggest that auditory modulation filtering affects the perception of both linguistic information and speaker individuality and there are dominant modulation regions for speech (bellow 5 Hz) and speaker recognition (bellow 8 Hz).

4. MTF-based speech signal processing

Figure 8 shows a general scheme for MTF-based speech signal processing as a function of modulation frequency (the dominant frequency in the temporal amplitude envelope) and the STI for predicting speech intelligibility derived from MTF [11]. The MTF is only of interest for the range of modulation frequency relevant for the speech signal; thus, this is for the range of fluctuation rhythms, which is determined by the spectrum of the temporal amplitude envelope of speech. Based on the MTF, we can know how much noise and reverberation affect reduction in the MTF, and we can then predict reduced speech intelligibility using STI derived from MTF. The approach of MTF-based temporal processing is aimed at restoring reduced MTF and then enhancing speech intelligibility using the restored MTF as shown in Fig. 8(c).

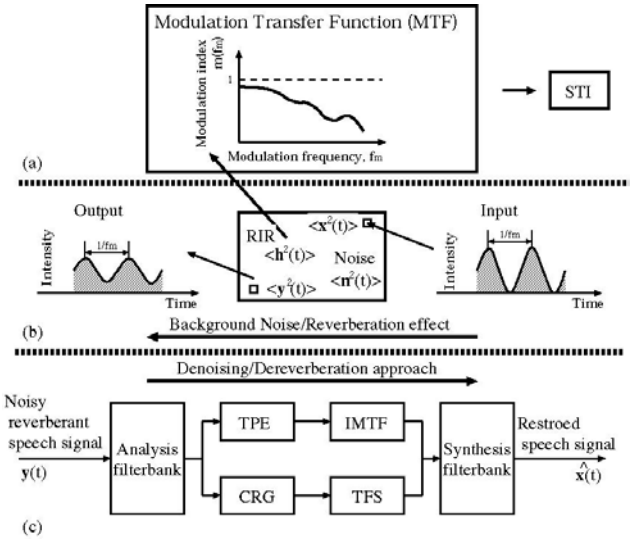

Figure 8. Processing on temporal amplitude modulation.

From the same approach in the noise-vocoder scheme, nonlinguistic information such as speaker individuality and emotion on temporal amplitude envelope may be able to be enhanced or controlled by using MTF-based temporal processing. This will be further works in our research.

5. Conclusion

In this study, the concept of amplitude modulation transfer was introduced to comprehensively study a role of auditory perception. The effect of controlling the highest modulation frequency of noise-vocoded speech on the recognition of words and speaker were investigated to clarity the modulation frequency bands related to the perception of linguistic and speaker individuality information. The highest modulation frequency of speech signal was controlled by low-pass filtering the temporal amplitude envelope of speech. The results of word intelligibility tests showed that linguistic information decreased when the highest modulation frequency was less than 5 Hz. This suggests that the shape of the temporal envelope of the moraic syllable structure is an important factor in recognizing linguistic information. The results of speaker identification experiment showed that the modulation components below 16 Hz should contribute to the perception of speaker individuality information. Different from the perception of linguistic information, higher variations of temporal envelope are important for speaker identification.

In future works, we will investigate whether nonlinguistic information on temporal amplitude envelope may be able to be controlled by using MTF-based temporal amplitude processing.

Acknowledgment

This work was supported by the Okawa Foundation for Information and Telecommunication. This study was partially supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and by the Secom Science and Technology Foundation. The author thanks to

References

[1] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. “Speech recognition with primarily temporal cues,” Science, vol. 270, pp. 303-304, 1995.

[2] Drullman, R. “Temporal envelope and fine structure cues for speech intelligibility,” J. Acoust. Soc. Am., vol. 97, no. 1, pp. 585–592, 1995.

[3] Dau, T. “Modeling auditory processing of amplitude modulation,” PhD thesis, Universität Oldenburg, 1996.

[4] Dau, T.and Kollmeier, B. “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers,” J. Acoust. Soc. Am., vol. 102, pp. 2892-2905, 1997.

[5] Houtgast, T. and Steeneken, H. J. M. “The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility,” Acustica, vol. 28, pp. 66–73, 1973.

[6] Atlas, L., Greenberg, S. and Hermansky, H. “The Modulation Spectrum and Its Application to Speech Science and Technology,” Interspeech2007, Tutorial, 2007.

[7] Arai, T., Pavel. M., Hermansky, H., and Avendano, C. “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” J. Acoust. Soc. Am., vol. 105, no. 5, pp. 2783-2791, 1999.

[8] Unoki, M., Irino, T., Glasberg, B, Moore, B. C. J., and Patterson, R. D., “Comparison of the roex and gammachirp filters as representations of the auditory filter,” J. Acoust. Soc. Am., vol. 120, no. 3, pp. 1474-1492, 2006.

[9] Nishino, Y., Miyauchi, R., and Unoki, M., “Study on Linguistic Information Contained in Temporal Amplitude Envelope of Japanese Speech Signals,” Proc. NCSP14, pp. 333-336, March 2014.

[10] Zhu, Z., Miyauchi, R., and Unoki, M., “Analysis of Speaker Individual Differences on Modulation Spectrum,” Proc. NCSP15, pp. 17-20, March 2015.

[11] Unoki, M., “Speech enhancement based on the concept of modulation transfer function,” Systemtheorie Signalverarbeitung Sprachtechnologie, Studient texte zur Sprachkommunikation Band 68, 109-116, TUD Press, 2013.