ESTIMATING NUMBER OF SPEAKERS BY THE MODULATION CHARACTERISTICS OF SPEECH

Takayuki Arai
(arai@sophia.ac.jp)
Dept. of Electrical and Electronics Eng., Sophia University, Tokyo, JAPAN

ABSTRACT

A method for estimating number of speakers of mixed speech signals was proposed. The algorithm was based on the modulation characteristics of speech, specifically that a single speech utterance typically has a distinct modulation pattern with a peak around 4-5 Hz. Having observed that the modulation peak decreases as number of speakers increases, our estimation algorithm used the region of the modulation frequency between 2 and 8 Hz. We obtained a novel parameter we called "equivalent number of speakers" to estimate the number of simultaneous speakers when speech signals contain multiple speakers.

1. INTRODUCTION

A speech signal is often modeled as an amplitude-modulated signal, where the resonator acts as a modulator and the sound source acts as a carrier. Usually, the frequency characteristics of the resonator vary as a function of time. Thus, the modulator dynamically changes in time, and therefore, it is important to consider the modulation characteristics when we describe speech signals. It has been reported by many researchers that the 4-5 Hz components of the modulation frequency are typically dominant when temporal dynamics of speech features are analyzed (Houtgast and Steeneken, 1985; Greenberg, 1995; Arai and Greenberg, 1997).

On the other hand, the human auditory system has band-limited characteristics for the temporal dynamics of speech, such that speech intelligibility declines once the 1-16 Hz modulation components of the speech signal are lost (Drullman et al., 1994; Arai et al., 1999). Thus, modulation characteristics are crucial in speech perception.

The modulation pattern is also useful when discriminating speech from other signals, such as noise and music (e.g., Karneback, 2001). In this study, we first analyze the modulation characteristics of mixed speech signals uttered by multiple speakers at a time. Second, we propose a novel parameter called "equivalent number of speakers" to estimate the number of speakers talking simultaneously.

2. CHANGE OF THE MODULATION SPECTRUM VS. NUMBER OF SPEAKERS

When a speaker is talking, the speech signal typically has a distinct modulation spectrum with a peak around 4-5 Hz. When more than one speaker is talking at the same time, the mixed speech signal has a more complex modulation pattern. In fact, each modulation pattern is out of phase relative to the others. As a result, the modulation spectrum becomes less distinct.

To see the change in shape of the modulation spectrum, we conducted a simple analysis experiment. First, eight TIMIT sentences were chosen randomly. Then, mixed speech signals were obtained by adding time signals after normalizing the signals with their root-mean-square values. All combinations were calculated; Table 1 shows the number of combinations versus the number of speakers.

Table 1: Number of combinations vs. number of speakers for the analysis experiment.

#speaker	1	2	3	4	5	6	7	8
#combinations	8	28	56	70	56	28	8	1

Figure 1 shows the spectrograms of mixed speech signals from one (a) to eight (h) speakers. As you can see from Fig. 1, the patterns are smeared as the number of mixtures increases.

The process used to obtain the eight lines of data in Fig. 2 is as follows:

Eight (TIMIT) sentences were used, s1 to s8.

For 1 speaker, s1 to s8 were used to compute the modulation spectrum for each sentence. Eight modulation spectra were obtained, the results were averaged, and one line was obtained.

For 2 speakers, all combinations were made: 1) s1+s2, 2) s1+s3, 3) s1+s4, 4) s1+s5, 5) s1+s6, 6) s1+s7, 7) s1+s8, 8) s2+s3, 9) s2+s4, 10) s2+s5, 11) s2+s6, 12)s2+s7, 13) s2+s8, 14) s3+s4, 15) s3+s5, 16) s3+s6, 17) s3+s7, 18) s3+s8, 19) s4+s5,
20) s4+s6, 21) s4+s7, 22) s4+s8, 23) s5+s6, 24) s5+s7, 25) s5+s8, 26) s6+s7, 27) s6+s8, 28) s7+s8. Twenty-eight modulation spectra were computed for all 28 mixtures (each mixture contains two-speaker's speech sounds), and averaged, and one line was obtained.

For 3 speakers, all 56 combinations were made. The 56 modulation spectra from all 56 mixtures were computed, averaged, and one line was obtained.

The process continued similarly for 4-8 speakers. For 4 speakers, 70 combinations were made and averaged and one line was obtained. For 5 speakers, 56 combinations were made and averaged and one line was obtained. For 6 speakers, 28 combinations were made and averaged and one line was obtained. For 7 speakers, 8 combinations were made and averaged and one line was obtained. For 8 speakers, only one combination was made, and one line was obtained.

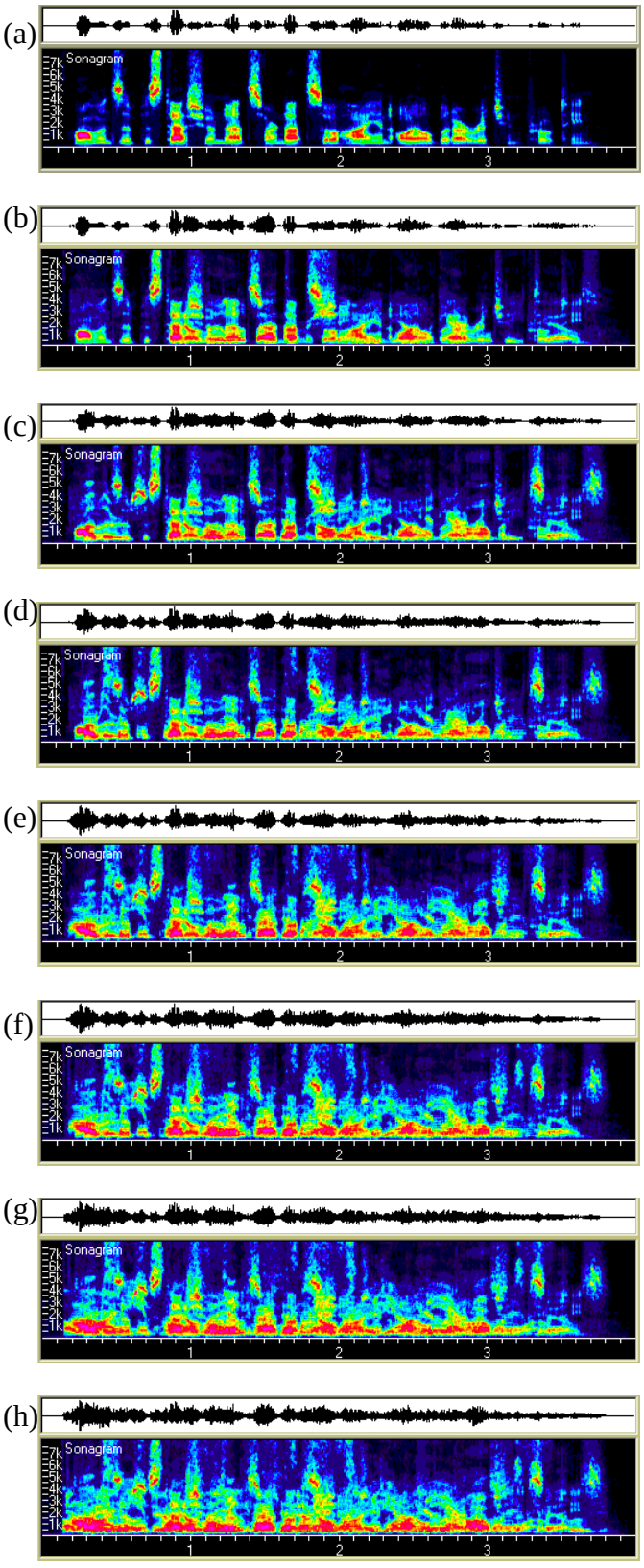

Fig. 1: Spectrograms of mixed speech from one (a) to eight (h) speakers.

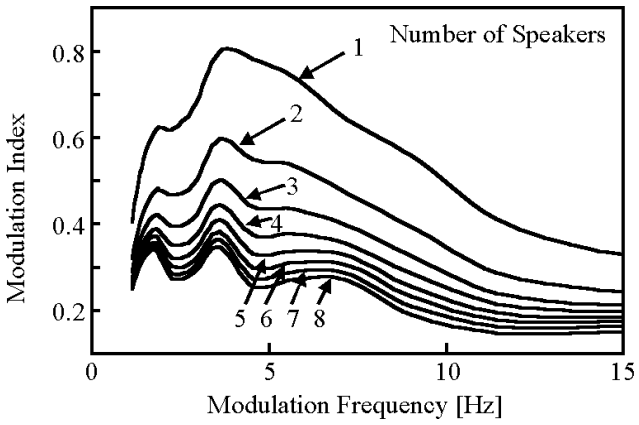

Fig. 2: Average modulation spectra of the eight TIMIT sentences. Each line corresponds to the average modulation spectrum for each number of speakers.

Figure 2 shows the average modulation spectra of the eight TIMIT sentences. To compute this figure, 500-2000 Hz region was used. This is because this frequency region has more energy, so that the modulation characteristics can be more stable (Greenberg and Arai, 1998). Each line corresponds to the average modulation spectrum for each number of speakers. The average was taken among all combinations for each mixture. This figure shows the change of the modulation spectrum as the number of speakers increases. Modulation indices in the range between 2 and 8 Hz of the modulation frequency decrease as a function of the number of speakers.

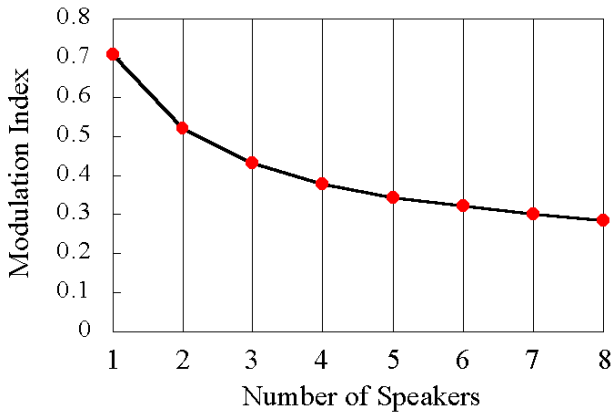

Fig. 3: Modulation index vs. number of speakers.

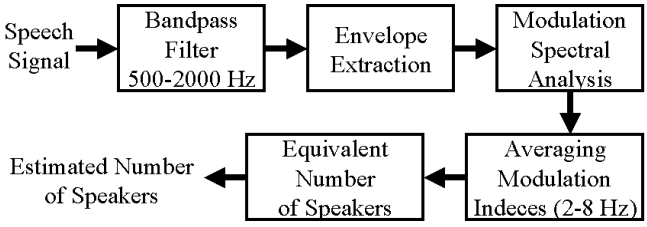

Fig. 4: Block diagram of the proposed estimation procedure of the number of speakers.

3. EQUIVALENT NUMBER OF SPEAKERS

In Fig. 2, we saw that the modulation indices between 2 and 8 Hz decreased as the number of speakers increased. This, in fact, suggests that we can compute a curve for estimating the number of speakers of unknown speech signals. In other words, we can define a curve of "equivalent number of speakers" and used it to estimate number of speakers by analyzing the average modulation indices in the range of 2-8 Hz of the modulation frequency.

Figure 3 shows the modulation index vs. the number of speakers. This curve was computed by averaging the modulation indices between 2-8 Hz of the modulation frequency and plotted as a function of the number of speakers. From this figure, we confirmed the declining tendency in an explicit way.

Figure 3 can be modeled based on the minimum-square-error criterion as follows:

$$MI = \frac{1}{N_s + 0.8697} + 0.1734, \qquad (1)$$

where MI is the modulation index and N_s is the number of speakers. The curve of the "equivalent number of speakers" is defined as the inverse function of Equation (1).

By using this curve, we are able to estimate the number of speakers in a mixed signal. Figure 4 shows the block diagram of this proposed estimation procedure. A speech signal was first input to the bandpass filter between 500-2000 Hz. Then, the envelope of the band-limited signal was computed. The modulation indices between the modulation frequencies of 2 and 8 Hz were computed. Finally, an estimation of the number of speakers was obtained by looking up the "equivalent number of speakers" curve with the averaged modulation indices. For example, let us suppose an average

modulation index of a signal is 0.5. The curve indicates that there can be two speakers talking at the same time, because the estimation from the curve is 2.2.

4. CONCLUSIONS

We proposed a novel method for estimating number of speakers of mixed speech signals. The algorithm was based on the modulation characteristics of speech. We first analyzed TIMIT sentences and observed that the modulation peak decreases as the number of speakers increases. Finally, we established the estimation algorithm by using the region of the modulation frequency between 2 and 8 Hz. The curve of "equivalent number of speakers" allowed us to estimate the number of simultaneous speakers when speech signals are mixed by multiple speakers.

ACKNOWLEDGEMENTS

I would like to thank to Prof. Tachibana of the Univ. of Tokyo, and the members of his lab., especially to Ami Aoki. I would also like to express my appreciation to Terri Lander for her useful comments.

REFERENCES

[1] T. Houtgast and H. J. Steeneken, "A review of the MTF concept in room acoustics and its use of estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, Vol. 77, pp. 1069-1077, 1985.

[2] S. Greenberg, "The ear have it: The auditory basis of speech perception," *Proc. of Int'l Congress of Phonetic Sciences*, Vol. 3, pp. 34-41, 1995.

[3] T. Arai and S. Greenberg, "The temporal properties of spoken Japanese are similar to those of English," *Proc. of Eurospeech*, Vol. 2, pp. 1011-1014, 1997.

[4] R. Drullman, J. M. Festen and R. Plomp, "Effect of temporal envelope smearing of speech reception," *J. Acoust. Soc. Am.*, Vol. 95, pp. 1053-1064, 1994.

[5] R. Drullman, J. M. Festen and R. Plomp, "Effect of Reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, Vol. 95, pp. 2670-2680, 1994.

[6] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, Vol. 105, pp. 2783-2791, 1999.

[7] S. Karneback, "Discrimination between speech and music based on a low frequency modulation feature," *Proc. of Eurospeech*, Vol. 3, pp. 1891-1894, 2001.

[8] S. Greenberg and T. Arai, "Speech intelligibility derived from exceedingly sparse spectral information," *Proc. of ICSLP*, Vol. 6, pp. 2803-2806, 1998.