FIFTH INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
DECEMBER 15-18, 1997
ADELAIDE, SOUTH AUSTRALIA

ACOUSTIC SIGNAL PROCESSING USING MULTI-WINDOWED STFT AND HARMONICS SIEVING

Takahiro Ohnishi[*], *Michiko Kazama*[**], *and* *Mikio Tohyama*[*]
大西 崇浩　風間 道子　東山 三樹夫
[*]Kogakuin University
[**] Acoustic Consultant

1. ABSTRACT

This article describes signal enhancement by estimating the fundamental frequency and harmonics of a nonstationary signal using Multi-Windowed Short-time Fourier Transform and Harmonics Sieving(MWSTFT-HS). The MWSTFT-HS which we propose here is based on the sinusoidal wave model and analyzes a signal with high frequency resolution by using multi-rectangular windows, each of which has a different window length. First, the authors confirmed that the fundamental frequencies and harmonics of a speech waveform can be estimated using MWSTFT-HS, and successfully discriminated the vowels from the consonants based on the strength of the harmonic structure. Next, using results of fundamental studies the authors explored the possibilities of separating sung or spoken sound from simultaneously recorded signals. This study confirmed that the sound of singers or talkers whose voices have different fundamental frequencies could be well separated or enhanced by tracking the temporally changing fundamental frequencies and harmonics. These investigations can be applied to signal enhancement under noisy conditions.

2. INTRODUCTION

Signal enhancement is an important technical area for acoustic signal processing and therefore various methods of signal enhancement have been investigated. Co-talker separation based on the cocktail party effect simulation[1] and a method using power spectrum peak-picking[2] have been proposed. We noticed from the results of those investigations that improving the frequency resolution of signal analysis could be a key technology for signal enhancement.

This article proposes a method for signal analysis with high frequency resolution[3][4][5] based on the sinusoidal signal model, and investigates signal

enhancement using 1-point microphone reception. If we can estimate the fundamental frequency of a signal and the harmonics, we can construct the signal waveform according to the sinusoidal model. We extend the conventional sinusoidal model[6] to include compound sinusoidal waves which are composed of different fundamental frequencies and harmonics[7]. The authors use this extended model to investigate separation of the sounds from singers and talkers. The results show that fundamental frequency estimation is a significant cue as long as the sound of the talker or singer have different fundamental frequencies and harmonics.

3. PRINCIPLE OF MWSTFT AND HARMONICS SIEVING

Figure 1 shows our fundamental frequency estimation procedure and harmonics sieving using MWSTFT. The frequency resolution of an STFT depends on the length of its short-time window. The MWSTFT we propose here performs high-resolution frequency analysis by using multi-rectangular windows. The procedure is as follows.

(1) MWSTFT gets an original signal using a rectangular window whose length is L frames. It again windows its L-length signal using rectangular windows whose lengths range from L to (L/2+1) windows.
(2) We obtain the DFTs by using the multi-rectangular windows with lengths from L to (L/2+1). Since each MWSTFT is able to analyze different frequency components, we can describe a signal with high resolution. The next part deals with harmonics sieving.
(3) Every spectral component obtained by MWSTFT is a candidate fundamental frequency. We take a summation of the power spectra $POW_N(l,p)$ for p harmonics of every candidate l in the i-th window where $N = L - i + 1$ (Fig. 1, Eq. (1)). Here every $POW_N(l,p)$ is normalized by the signal power of the i-th window. This is our harmonics sieving function.
(4) We find the STFT that best matches the original waveform when $POW_N(l,p)$ is maximized.
(5) Next we subtract the time wave components that are reconstructed from the fundamental frequency and harmonics estimated in the STFT that best matches the original waveform (Fig. 1, Eq. (2)). Here the time wave component is extended into the entire time frame L beyond the best-match window.
(6) We repeat these processes until we have extracted all the fundamental frequencies and harmonic components included in the original signal $x(n)$.

4. ESTIMATION OF FUNDAMENTAL FREQUENCY

Figure 2 illustrates distributions for the maxima of POW_N and the estimated fundamental frequencies. Figure 2(a) shows the estimated results for a compound sine wave of fundamental frequency 500 Hz and three harmonics. The frame length is 16 ms. The equation for POW_N is given by

$$POW_N = 20\log_{10}\left(\sum_{m=1}^{p}\left|X_N(ml)^2\right| / \sum_{l=1}^{(N-1)/2}\left|X_N(l)^2\right|\right)$$

The signal has 31 frames and the fundamental frequencies were estimated in every frame. All the plots of the estimated results are located at one point of (maxima of POW_N = 0 dB, 500 Hz). We can see that MWSTFT-HS exactly estimates the fundamental frequency for a compound sine wave of a harmonic structure. Figure 2 (b) is the distribution of estimated fundamental frequencies for a random noise. These plots are scattered irregularly since white noise has no harmonic structure.

5. ESTIMATION OF FUNDAMENTAL FREQUENCY OF A SIGNAL

Figure 3 shows distributions for the maxima of POW_N and the estimated fundamental frequencies of a speech signal [shi]. Figure 3(a) is a time waveform of the speech signal. Estimated results in the frames within the first 200 ms are shown by $\times$, and $\bigcirc$ represent the results obtained in the frames during the last 200 ms. The first 200 ms interval mostly contains the consonants, while the last 200 (ms) interval shows the vowel parts. The plotted $\bigcirc$ values look similar to the distribution in Fig. 2(a), while the plotted $\times$ values are, like the $\times$ values in Fig. 2(b), randomly distributed. From these distribution analysis of the POW_{NMax}, we can confirm that the vowel parts are discriminated from the consonant parts by MWSTFT-HS.

6. FUNDAMENTAL FREQUENCY ESTIMATION FOR SINGING VOICES

Estimation of composed melody lines is quite important for automatic musical score writing. As signal composed of different fundamental frequencies (f_0, f_1, f_2, ...) is written as

$$x(t) = \sum_{l=1}^{L} A_l e^{j2\pi l f_0 t} + \sum_{m=1}^{L} B_m e^{j2\pi m f_1 t} + \cdots + \sum_{n=1}^{L} C_n e^{j2\pi n f_{\cdot} t} \quad .$$

A chorus composed of different melody lines can be modeled as the signal described above. Figure 4 illustrates the melody lines - including vibratos - extracted from the chorus using the MWSTFT-HS. The fundamental frequencies are estimated every 32 ms and those fundamental frequencies are arranged in a high frequency order. We can see the melody lines are written almost correctly in all the parts. Every melody line can be separately synthesized using the fundamental frequency and harmonics.

Figure 5 illustrates the result obtained in a similar manner using conventional STFT instead of MWSTFT-HS. The fundamental frequencies are not correctly estimated including the vibratos, since the frequency resolution of STFT is lower than that predicted by the MWSTFT-HS.

7. SEPARATION OF MALE AND FEMALE VOICES

We try using MWSTFT-HS to perform two-talker separation as well as separating singing voices. First, the fundamental frequencies of that male and female voices are estimated in short frames using MWSTFT-HS. Figure 6 shows that the histogram of

estimated fundamental frequencies of male voices is concentrated around 130 - 160 Hz, while the fundamental frequencies of female voices are mainly observed near 200 Hz. Suppose that our target signal is the female voice. If the estimated fundamental frequency is lower than 160 Hz, we decide that the frame is composed of mainly the male voice. On the contrary, if the estimated fundamental frequency is higher than 160 Hz, we decide that the frame is for the female voice.

Figure 7 shows the separation experiment results when the energy ratio is between male and female voices about 5 dB. If we compare the original female talker's voice shown in Fig. 7(b-1) as the separated one shown by Fig. 7(c-1), we can see that the female voice could be detected from the mixed voice recording. But there are frames where the male and female voices are still mixed up, particularly when low fundamental frequencies are observed.

8. CONCLUSION

This article has described signal enhancement using MWSTFT-HS. The authors have confirmed by computer simulation experiments that the proposed method has great potential for signal enhancement. Fundamental frequency estimation of a speech signal, and discrimination of vowels and consonants could be possible since the MWSTFT-HS method provided a super resolution for fundamental frequency analysis even in a short time interval. The vocal melody lines composed of soprano, mezzo-soprano, and alto were dramatically well separated by tracking the fundamental frequencies and the harmonics. Separation of the voices of two speakers from one microphone recording using MWSTFT-HS was promising, however improving the voice quality and separation in the parts where fundamental frequencies of the two speakers' voices overlapped are problems to be solved in the future. This study is partly supported by the Information - technology Promotion Agency, Japan (IPA) for creative software development projects '96 .

REFERENCES

1) Y. Cao, J. Audio Eng. Soc., 44(12), pp.1084-1096 (1996)
2) T. Houtgast, Private Communication (1996)
3) J. C. Brown, J. Acoust. Soc. Am., 94(2), pp.662-667 (1993)
4) W. J. Pielemeier and G. H. Wakefield, J. Acoust. Soc. Am. 99(4), pp.2382- 2396 (1996)
5) R. C. Maher, J. Acoust. Soc. Am., 95(4), pp. 2254-2263 (1994)
6) T. F. Quatieri al., ASSP, 34(6), pp. 1449-1464 (1996)
7) M. A. Cohen, S. Grossberg, and L. L. Wyse, J. Acoust. Soc. Am., 98(2), pp. 862-879 (1995)

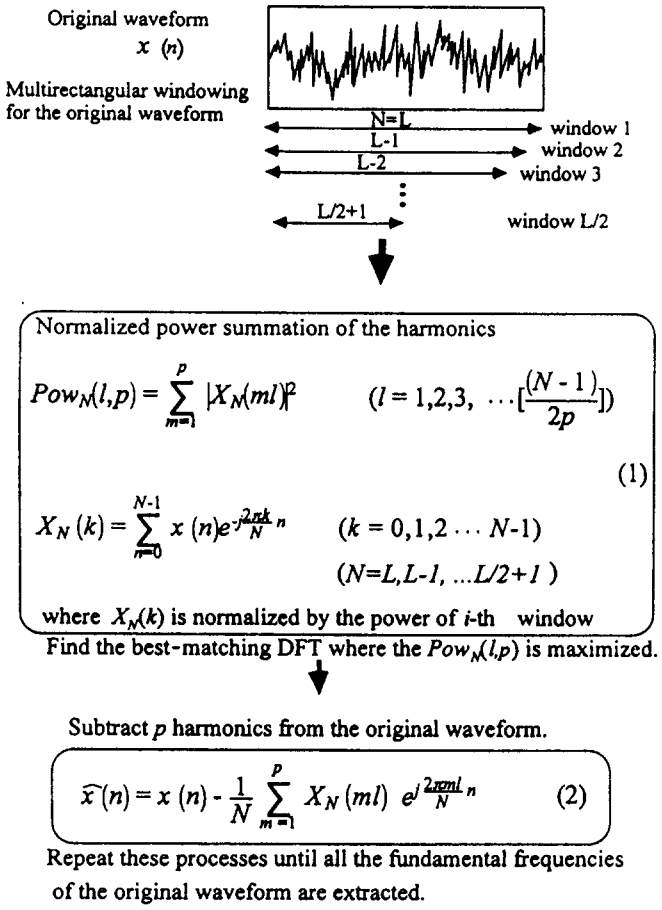

Figure 1. Principle of MWSTFT-HS.

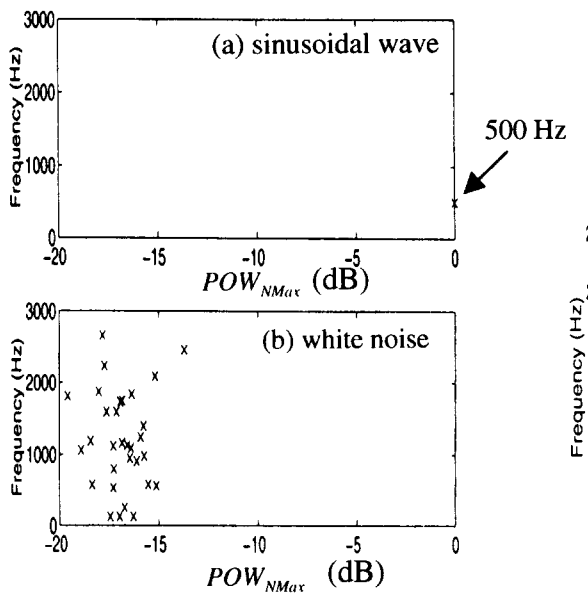

Figure 2 Distribution POW_{NMax} and estimation of fundamental frequencies

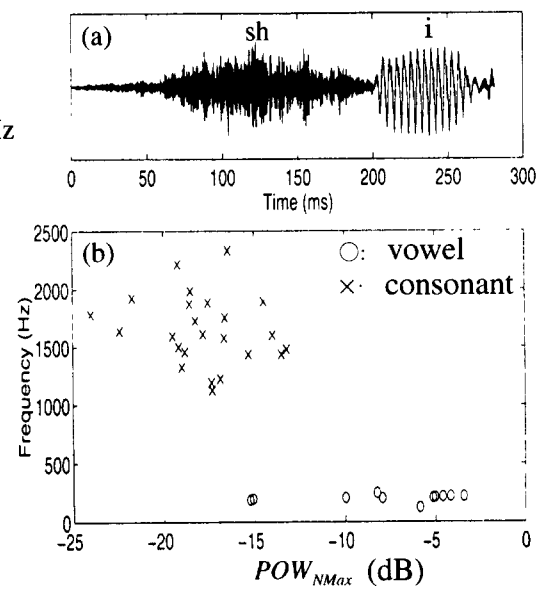

Figure 3 Time waveform of [shi](a), POW_{NMax} and fundamental frequencies(b)

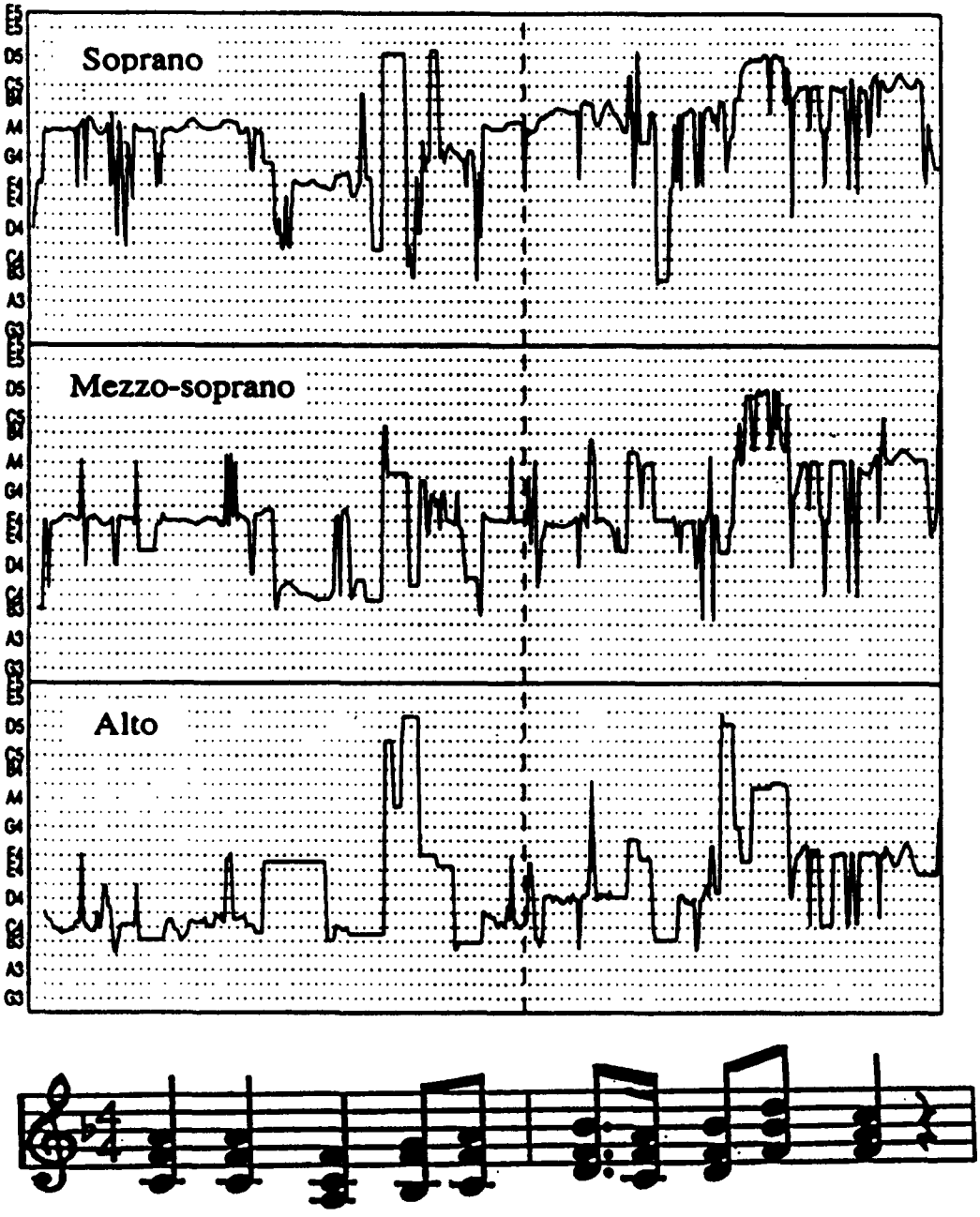

Figure 4 Fundamental frequencies estimated using MWSTFT-HS for female chorus (My Old Kentucky Home)

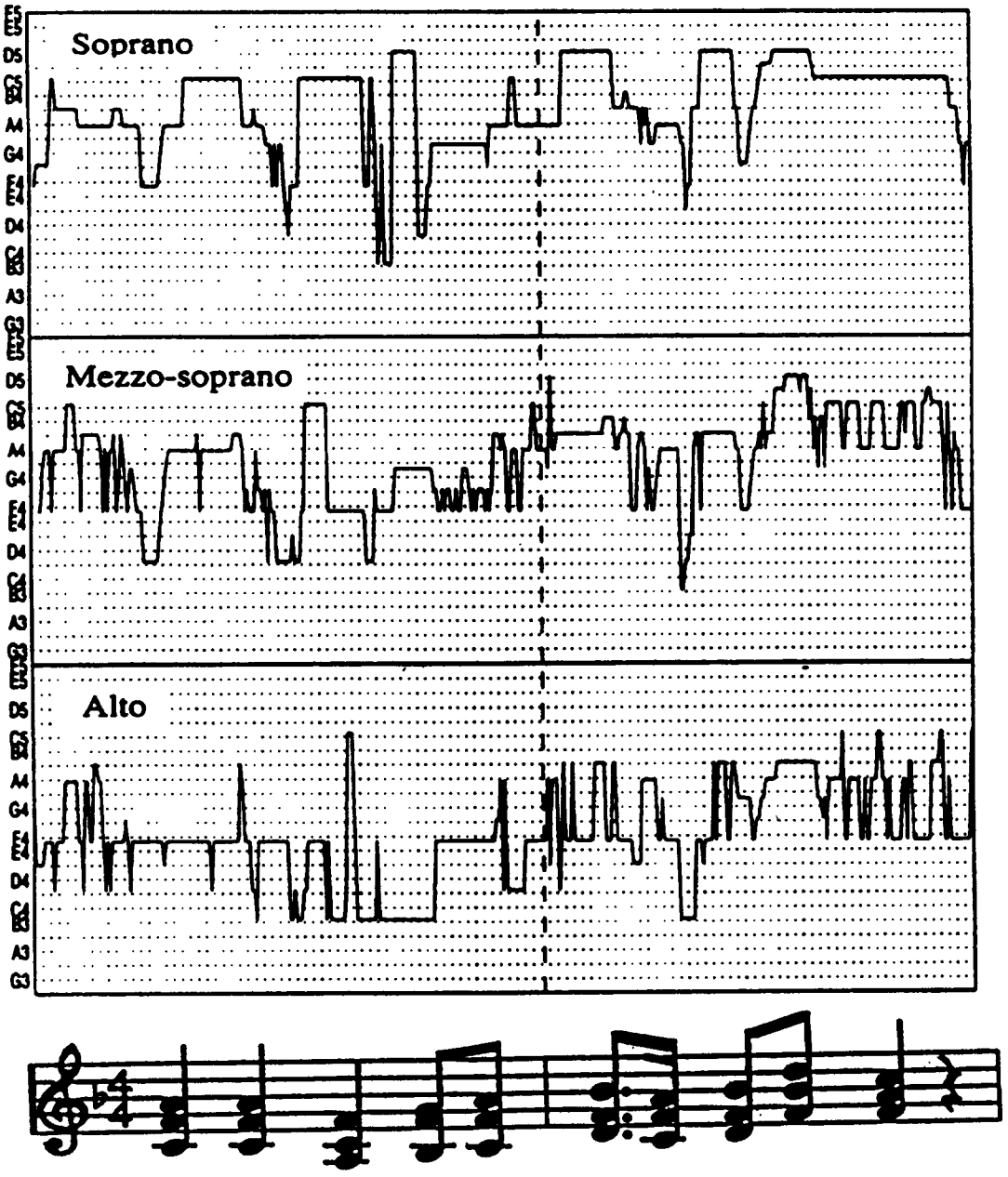

Figure 5 Fundamental frequencies estimated using STFT for chorus (My Old Kentucky Home).

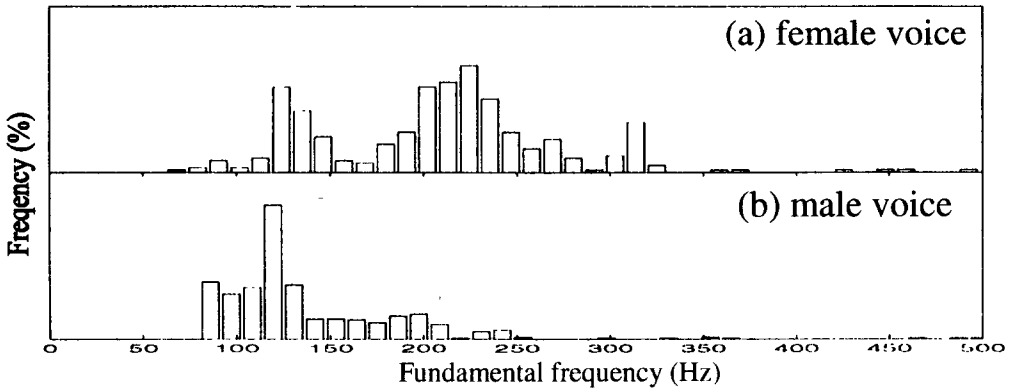

Figure 6 Histogram of fundamental frequencies estimated using MWSTFT-HS

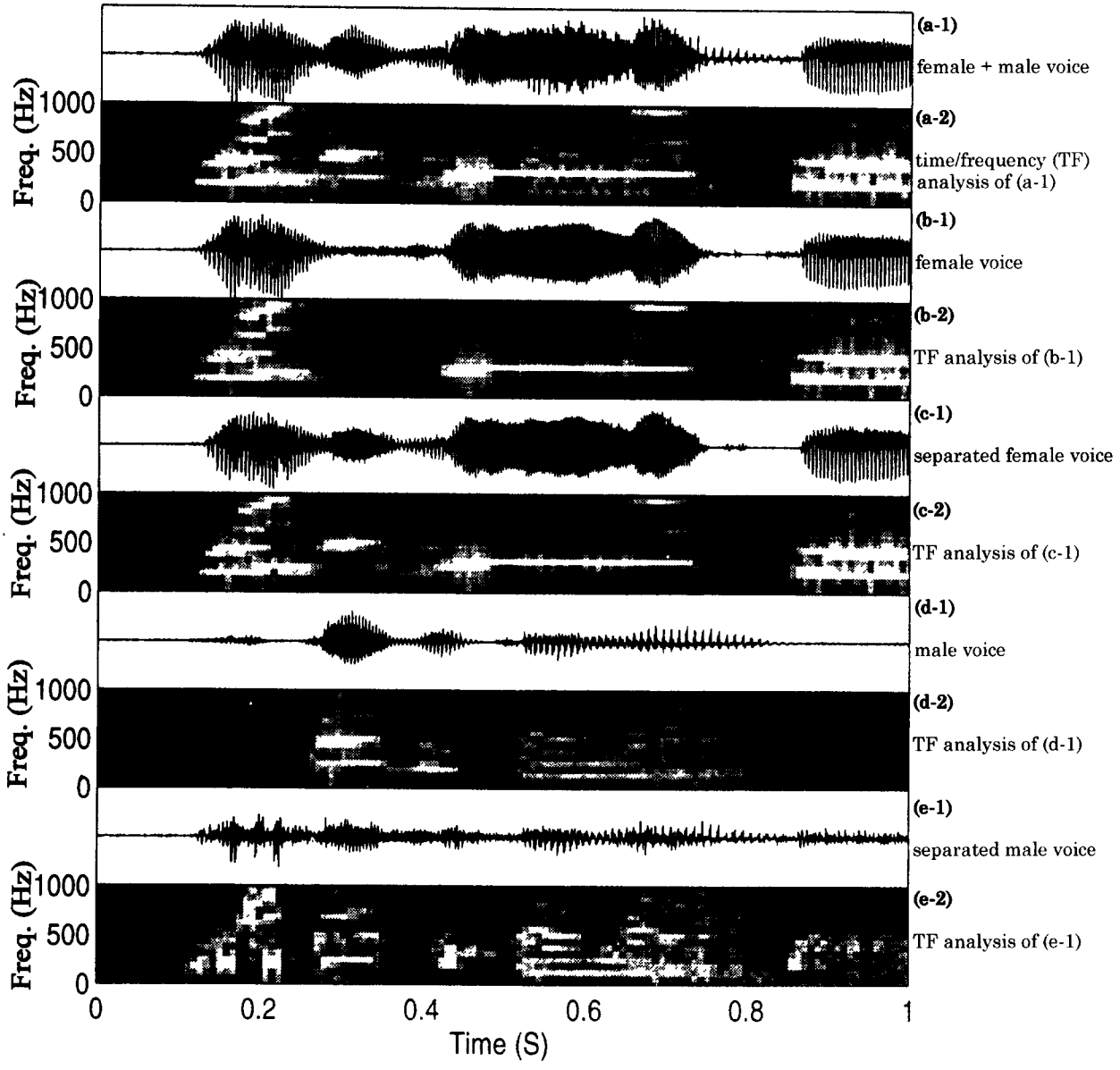

Figure 7 Separation of two – talkers' spoken sentences