

SPEECH SEGREGATION BASED ON FUNDAMENTAL EVENT INFORMATION USING AN AUDITORY VOCODER

Toshio Irino^{*}, Roy D. Patterson^{**}, and Hideki Kawahara^{*+}

^{*} Faculty of Systems Engineering, Wakayama University, Japan.

^{**} Centre for the Neural Basis of Hearing, Cambridge University, UK, ⁺ ATR, Japan.

irino@sys.wakayama-u.ac.jp; roy.patterson@mrc-cbu.cam.ac.uk; kawahara@sys.wakayama-u.ac.jp

ABSTRACT

We present a new auditory method to segregate concurrent speech sounds. The system is based on an auditory vocoder developed to resynthesize speech from an auditory Mellin representation using the vocoder STRAIGHT. The auditory representation preserves fine temporal information, unlike conventional window-based processing, and this makes it possible to segregate speech sources with an event synchronous procedure. We developed a method to convert fundamental frequency information to estimate glottal pulse times so as to facilitate robust extraction of the target speech. The results show that the segregation is good even when the SNR is 0 dB; the extracted target speech was a little distorted but entirely intelligible, whereas the distracter speech was reduced to a non-speech sound that was not perceptually disturbing. So, this auditory vocoder has potential for speech enhancement in applications such as hearing aids.

1. INTRODUCTION

Speech segregation is an important topic for speech signal processing. Many systems have been proposed since the 1970's [1,2]. Most of the systems are based on the short-time Fourier transform and segregate sounds by extracting harmonic components located at integer multiples of the fundamental frequency. It is, however, difficult to extract truly harmonic components when the fundamental frequency is disturbed by concurrent noise; the error increases in proportion with the harmonic number.

We have developed an alternative method for speech segregation based on the Auditory Image Model (AIM) and a scheme of event-synchronous processing. AIM was developed to provide a reasonable representation of the "auditory image" we perceive in response to sounds [3]. We have also developed an "auditory vocoder" [4,5] for resynthesizing speech from the auditory image using an existing, high-quality vocoder, STRAIGHT [6]. The auditory representation preserves fine temporal information, unlike conventional window-based processing, and this makes it possible to do synchronous speech segregation.

In a previous paper, we presented a method to segregate a target speaker from a mixture of concurrent speech using an event-synchronous version of the auditory vocoder. We also demonstrated the potential of the method in low SNR conditions when the event-time information of the target speech is known in advance. It is, however, more difficult

to extract precise event times than the fundamental frequency (F0) particularly in low SNR conditions. So, we have developed a method to convert the F0 to event times. In this paper, we describe an event-synchronous version of the auditory vocoder, a method for converting F0 to event times, and the potential of the system for improving speech segregation.

2. AUDITORY VOCODER

The system has four components (Fig. 1): There is a robust F0 estimator [e.g., 8], the STRAIGHT vocoder [6] and the Auditory Image Model (AIM) [3] which together produce the Mellin Image [7], and there is a mapping block to link the auditory model and the vocoder. In this section, we concentrate on the synchronization procedure and a method for converting F0 to event times; the rest of the system has been described previously [4, 5].

2.1. Event based Auditory Image Model

AIM performs its spectral analysis with a gammatone filterbank on a quasi-logarithmic (ERB) frequency scale. The output is half-wave rectified and logarithmically compressed. Then, adaptive thresholding is applied in each channel to produce a simple form of Neural Activity Pattern (NAP). The NAP is converted into a Stabilized Auditory Image (SAI) using a strobe mechanism controlled by an event detector. Basically, it calculates the times between neural pulses in the auditory nerve and constructs an array of time-interval histograms, one for each channel of the filterbank.

2.1.1. Event detection

An event-detection mechanism was introduced to locate glottal pulses accurately for use as strobe signals. The upper panel of Fig. 2 shows about 30 ms of the NAP of a male vowel. The abscissa is time in ms; the ordinate is the center frequency of the gammatone auditory filter in kHz. The mechanism identifies the time interval associated with the repetition of the neural pattern and the interval is the fundamental period of the speech at that point. Since the NAP is produced by convolution of the speech signal with the auditory filter and the group delay of the auditory filter varies with center frequency, it is necessary to compensate for group delay across channels when estimating event timing in the speech sound. The middle panel shows the NAP after group delay compensation; the operation aligns the responses of the filters in time to each glottal pulse. The solid line in the bottom panel shows the temporal profile derived by summing across channels in the compensated NAP in the region below 1.5 kHz. The

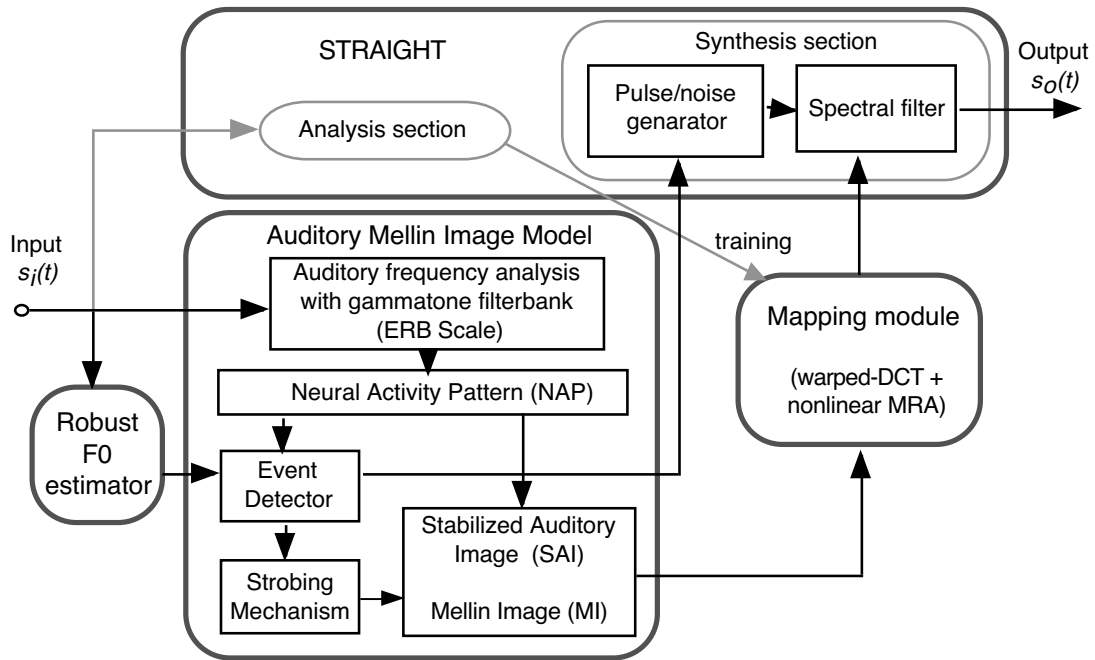


Figure 1. Configuration of auditory vocoder: The gray arrows show the processing path used for parameter estimation. Once the parameters are fixed, the path with the black arrows is used for resynthesizing speech signals.

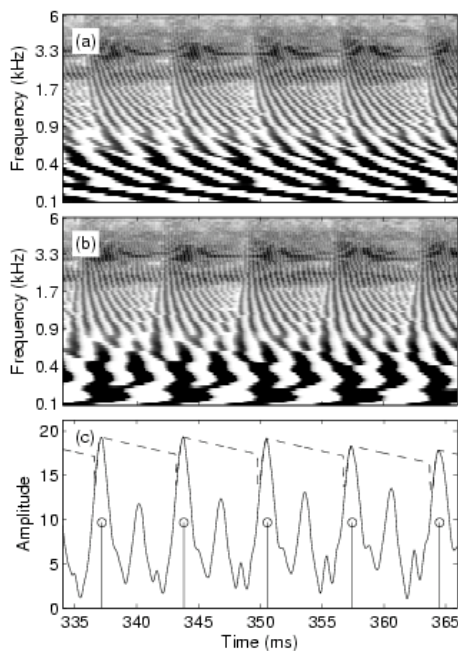


Figure 2. Auditory event detection for clean speech

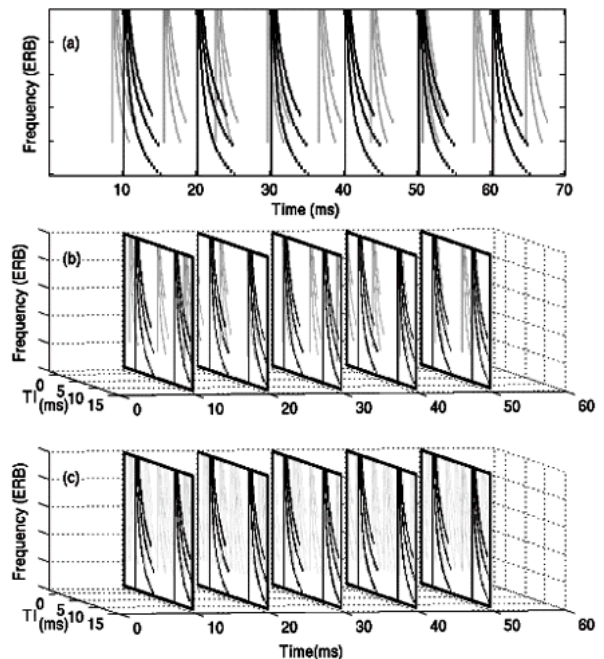


Figure 3. Event synchronous strobes for concurrent speech

peaks corresponding to glottal pulses are readily apparent in this representation.

We extracted peaks locally using an algorithm similar to adaptive thresholding [3]. The dotted line is the threshold used to identify peaks, indicated by circles and vertical lines. After a peak is found, the threshold decreases gradually to locate the next peak. We also introduced a form of prediction to make the peak detection robust. The threshold is reduced by a certain ratio when

the detector does not find activity at the expected period, defined as the median of recent periods. This is indicated by the sudden drop in the threshold.

Tests with synthetic sound confirmed that this algorithm works sufficiently well when the input is clean speech. It is, however, difficult to apply this method under noisy conditions particularly when the SNR is low. So, we enlisted the F0 information to improve event time estimation. The method is described in section 2.1.3. In

the next section, we describe a principle for speech segregation based on event synchronous strobing.

2.1.2. *Event-synchronous, strobed temporal integration*

Figure 3a shows a schematic plot of a NAP after group delay compensation for a segment of speech with concurrent vowels. The target speech with a 10-ms glottal period is converted into the black activity pattern, while the background speech with a 7-ms period is converted into the gray pattern. For every target event time (every 10 ms), the NAP is converted into a two-dimensional auditory image (AI) as shown in Fig. 3b. The horizontal axis of the AI is time-interval (TI) from the event time; the vertical axis is the same as the NAP. As shown in Fig. 3b, the activity pattern for the target speech always occurs at the same time-interval, whereas that for the background speech changes position. So, we get a "stabilized" version of the target speech pattern in this "instantaneous" auditory image (AI).

Temporal integration is performed by weighted averaging of successive frames of the instantaneous AI as shown in Fig. 3c, and this enhances the pattern for the target speech relative to that for the background speech. In this way, the fine structure of the target pattern is preserved and stabilized when the event detector captures the glottal event of the target speech correctly. The weighted averaging is essentially equivalent to conventional STI when the weighting function is a ramped exponent with a fixed half-life. In the following experiment, we used a hanning window spanning five successive SAI frames, although tests indicated that the window shape does not have a large effect as long as the window length is correct.

2.1.3. *A robust F0 estimator for event detection*

It is essential to detect the exact event times to make event-synchronous strobing work properly, and when the SNR is low, the precision of the event detector described in section 2.1.1 is not sufficient. It is easier to estimate fundamental frequency, F0, than event times for a given SNR. The latest methods for F0 estimation are robust in low SNR conditions and can provide accuracy that F0 values are correctly estimated within 5 % error for 80 % of voiced segments in babble noise at SNRs as low as 5 dB [8]. So, we developed a method of enhancing event detector with F0 information.

First, candidate event times are calculated using the event detection mechanism described in section 2.1.1. The half-life of the adaptive threshold is reduced to avoid missing events in target speech. The procedure extracts events for both the target and background speech. Then we produce a temporal sequence of the pulses located at the candidate times.

For every frame step (e.g., 3 ms), the value of F0 of the target speech in that frame is converted into a temporal function consisting of a triplet of gaussian functions with a periodicity of $1/F_0$. The triplet function is, then, cross-correlated with the event pulse sequence to find the best matching lag time. At this best point, the temporal sequence and the triplet of gaussians are multiplied together so as to derive a value similar to a likelihood for

each event. The likelihood-like values are accumulated and applied to the thresholding with an arbitrary value to derive estimates of the event times for the target speech signal.

2.2. Resynthesis of speech sounds

We can now produce an auditory image in which the pattern of the target speech is enhanced. It only remains to resynthesize the target speech from their auditory representations using the auditory vocoder and the mapping module (Fig. 1). We describe the method briefly here; see [5] for details. Each frame of the 2-D stabilized auditory image (SAI) is converted into a Mellin Image (MI), and then the MI is converted into a spectral distribution using a warped-DCT. The mapping function between the MI and the warped-DCT is constructed using a non-linear multiple regression analysis (MRA). Fortunately, it is possible to determine the parameter values for the non-linear MRA in advance, using clean speech sounds and the analysis section of STRAIGHT as indicated in Fig. 1 gray arrows. The speech sounds are, then, reproduced using a spectral filter excited with the pulse/noise generator shown in Fig. 1 in synthesis section of STRAIGHT. The pulse/noise generator is controlled by the robust F0 estimator in the left box.

3. EXPERIMENTS

It is difficult to estimate the exact F0 of the target speech when the SNR is low. This is, however, a common problem for any speech segregation system based on fundamental frequency extraction. So, in this section we concentrate on showing the potential of the event synchronous method for speech segregation when F0 values are extracted precisely from the isolated target speech.

3.1. Data and conditions

We used male (MHT) and female (FTK) speech from an ATR database of 503 sentences to evaluate the method. The sampling rate for STRAIGHT and the warped-DCT was set to 12 kHz to match the frequency range of the auditory filterbank (0.1-6 kHz). The mapping parameters were estimated using about 16 sentences (8 male and 8 female). The best nonlinear MRA parameters were determined after several runs of resynthesis using a subset of the speech data.

3.2. Performance of simple analysis/synthesis system

There was little difference between the original speech sound and the resynthesized sound when listening through loudspeaker. So the event-synchronous auditory vocoder appears to work properly for a single speaker in quiet. This version works for various applications including speech segregation although it would be desirable to improve the quality further.

3.3. Performance of speech segregation

Figures 4a and 4b show the original, isolated target speech and its spectrogram. The target sound was mixed with a speech distracter at an SNR of 0 dB to produce the wave in Fig. 4c. At this level, a listener needs to concentrate to hear the target speech in the distracter. The

spectrogram in Fig. 4d shows that there are many voiced segments from both of target and distracter speech sounds. The extracted target speech (Fig. 4e) was distorted but entirely intelligible, whereas the distracter speech was converted into a non-speech sound which greatly reduces the disturbance that it would otherwise produce. So it is easier to listen to the target sound as an extracted sound than in the original mix with the distracter. The combination is perceived as a single stream of speech with an odd noise background. This contrast appears in the spectrogram in Fig. 4f. The voiced segments of the distracter speech sound (e.g., in the region 1.0~1.2 s, 1.5~1.9 s, and 3.8~4.2 s) were converted into unvoiced segments as the spectra do not have any harmonic structure. Although the distracter is still apparent in the log-spectral representation, the change in excitation improves the perceptual distinction between the original and distracter speech. We think the performance could be improved further by applying noise reduction techniques at the auditory image stage.

3.4. Effect of F0 estimation error

In the previous section, it was assumed that the F0 values were estimated precisely. This is not always the case in realistic environments. The event times are estimated from F0 values as described in section 2.1.3. We need to examine the robustness of the method when there is a mismatch between the periodicity values from the NAP and $1/F_0$. So, we synthesized sounds modified F0 values that were 5, 10, or 20 % above the original F0 value. Informal listening indicates there is almost no effect for +5 %; indeed, it is difficult to detect the difference. There is a small effect for +10 %; some phonemes are disrupted. There is large effect for +20%; the resynthesized sound has almost no phonetic information. This indicates that the method is reasonably robust to F0 estimation error. We expect it is better than that for conventional methods based on harmonic component selection in the frequency domain, since the errors are not proportional to harmonic number. It would, however, be necessary to compare the methods quantitatively in a future study.

4. CONCLUSIONS

We have described a new method for speech segregation using an event-synchronous auditory vocoder. It enhances the intelligibility of the target speaker in the presence of concurrent background speech when the F0 value of the target speech is estimated precisely. We also found that the error in F0 estimation is not a serious problem provided it is less than about 5 %. The auditory vocoder should be useful for speech enhancement applications such as hearing aids.

Acknowledgments This work was partially supported by a project grant from faculty of systems engineering of Wakayama University.

REFERENCES

- [1] Parsons, T.W., "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, 60, pp.911-918, 1976.
- [2] Lim, J.S. Oppenheim, A.V. Braid, L.D., "Evaluation of an adaptive comb filtering method for enhancing speech degraded

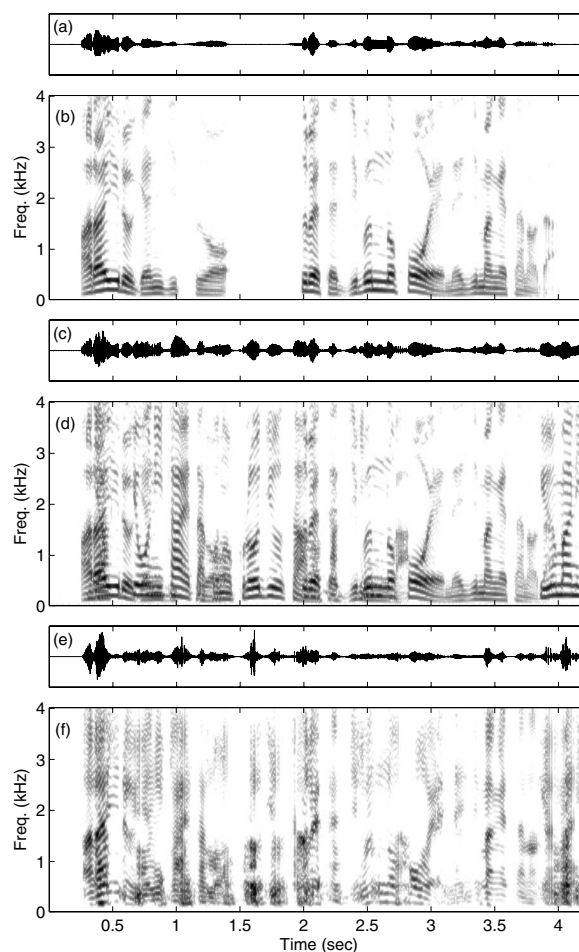


Figure 4 Speech waveforms and spectrograms: (a)(b) original clean speech, (c)(d) concurrent speech signal, (e)(f) after processing by the event-synchronous procedure.

by white noise addition," *IEEE, Trans. ASSP*, ASSP-26, 354-358, 1978.

- [3] Patterson, R.D., Allerhand, M. and Giguere, C. "Time-domain modelling of peripheral auditory processing: a modular architecture and a software platform", *J. Acoust. Soc. Am.*, 98, 1890-1894, 1995.
<http://www.mrc-cbu.cam.ac.uk/cnbh/>
- [4] Irino, T., Patterson, R.D., Kawahara, H. "Speech segregation using event synchronous auditory vocoder," in *Proc. IEEE ICASSP 2003, Hong Kong, Apr., 2003*.
- [5] Irino, T., Patterson, R.D., Kawahara, H., "An auditory vocoder resynthesis of speech from an auditory Mellin representation," *EAA-SEA-ASJ, Forum Acusticum Sevilla 2002*, HEA-02-005-IP, Sevilla, Spain, Sept., 2002.
- [6] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A. "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, 27, pp.187-207, 1999.
- [7] Irino, T. and Patterson, R.D. "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The Stabilised wavelet-Mellin transform," *Speech Communication*, 36, 181-202, 2002.
- [8] Nakatani, T. and Irino, T. "Robust fundamental frequency estimation against background noise and spectral distortion," *ICSLP 2002, 1733-1736, Denver, Colorado, USA, 16-20, Sept., 2002*.