# A COMPUTATIONAL MODEL OF SOUND STREAM SEGREGATION WITH MULTI-AGENT PARADIGM

*Tomohiro Nakatani, Takeshi Kawabata, and Hiroshi G. Okuno*
NTT Basic Research Laboratories
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-01 JAPAN
*nakatani@nuesun.ntt.jp, kaw@siva.ntt.jp, okuno@nuesun.ntt.jp*

## ABSTRACT

This paper presents a new computation model for sound stream segregation based on a multi-agent paradigm. Sound streams are thought to play a key role in auditory scene analysis, which provides a general framework for auditory research including voiced speech and music. Each agent is dynamically allocated to a sound stream, and it segregates the stream by focusing on consistent attributes. Agents interact with each other to resolve stream interference. In this paper, we design agents to segregate harmonic streams and a noise stream. The presented system can segregate all the streams from a mixture of a male and a female voiced speech and a background non-harmonic noise.

## 1. INTRODUCTION

Sound stream segregation from a sound mixture is a primary processing for understanding auditory events in the real-world (i.e., *Auditory Scene Analysis* [1]). Since there are too many sounds to identify all the auditory events, it is necessary to focus attention on certain streams. This phenomenon is known as the cocktail party effect. One of our motivations is to model the cocktail party effect and achieve it by computer. Since various kinds of sounds appear and terminate dynamically in the real-world, an adaptive mechanism is required for sound stream segregation. This paper presents a new approach toward stream segregation based on a multi-agent paradigm. Each agent is dynamically allocated to a sound stream, and traces the stream exclusively by using consistent attributes. Since sound streams interfere with each other in a sound mixture, agents interact with each other to restore missing sound streams using interference among environmental streams. As a whole, the system adaptively segregates individual sound streams. This paper reports on the design and evaluation of agents that segregate harmonic sounds and a background noise.

## 2. SOUND STREAM SEGREGATION

The following tasks are important for sound stream segregation, 1) to find a stream based on its consistent attributes, and 2) to restore the sound structure. Many techniques to segregate sound streams have been presented. Most use the auditory model for primary processing [2, 3]. The auditory map model proposed by Brown is a well-structured computation model [2]. That model first extracts all the sound features and makes several auditory maps. Then, it segregates streams by integrating sound features. Since the integration process becomes complicated when treating a real, complex sound mixture, blackboard architecture is now being used to simplify this integration process [4]. The auditory map model has some practical limitations. It is an off-line algorithm in the sense that any part of the input is available to the algorithm at any time. In addition, auditory maps carry richer information than is needed to segregate some sounds.

Using interfering signal cancellation, some specific sounds can be eliminated from the input [5, 6], and the processing of the other sounds then becomes easy. For example, Residual Interfering Signal Canceler (RISC) can effectively segregate some sound streams by using specific sound attributes [6]. Therefore, RISC represents a simple and powerful way to resolve stream interference when the attributes of each signal are known. However, the mechanism by which RISC architecture treats the number of sound sources and copes with different attributes of sounds has not been considered in any depth. These mechanisms are important for sound stream segregation.

We introduce a multi-agent paradigm into sound stream segregation. The system adapts to changes in the number of input streams by dynamically generating and terminating agents to treat individual streams. The sound streams are restored through the interaction among the agents. We have concentrated on frequency structure of sound streams for sound restoration. Al-

0-7803-2431-5/95 $4.00 © 1995 IEEE

though temporal structure of sound streams is also important, we do not deal with this issue in this paper [7].

## 3. DESIGN OF AGENTS

### 3.1. System structure

The stream segregation system must 1) determine that streams occur, 2) trace the streams, 3) detect that the streams have ended, and 4) resolve any interference between simultaneous streams. We have designed a multi-agent system for segregating streams (Figure 1). The system consists of two types of agents, *watchers* and *tracers*. A watcher finds new streams and generates a tracer. When two or more watchers detect a new stream simultaneously, they compete with each other. Some watchers are inactivated at that time. The set of watchers is called a *generator* since it generates tracers dynamically. Each tracer extracts a sound stream until it detects the end of the stream. Tracers also make subtract signals to remove their streams from input to other agents. This signal cancellation helps each agent to resolve stream interference, that is, it helps each watcher to avoid detection of redundant streams and helps each tracer to ignore other streams.

Sound is put into the system at each time frame (30-ms frame period, with a hamming window). Each agent receives a residual signal called a "*residual input*" after all the subtract signals are subtracted from the input. Each watcher uses the residual input to detect the new streams, and each tracer uses it to extract its stream. The system uses two types of subtract signals: waveform and spectrum signals. The waveform signal is subtracted from input, while the spectrum signal is sent directly to each agent.

We have already described the framework of this system [8]. We call this the *old system*. In the old system, redundant tracers were occasionally generated due to imperfect subtract signals, or imperfect exclusive sound allocation. In addition, the old system had only a crude mechanism to test if the consistency of each stream held. The main cause of these problems was thought to be that each agent checked only one input frame at a time. In the version reported in this paper, we provide several frames of input sound spectrum called "*spectrum input*" to the system at a time, and introduce a mechanism to solve the old system's problems. Some delay is added between the spectrum input and the residual input so that each agent can access the spectrum input prior to the residual input. Each agent uses the spectrum input to check how the sound consistency holds on it before using the residual signal to extract sound features. Each agent first estimates its stream by using the spectrum input and checks this consistency. At this time, agents communicate with each other to exclusively allocate sound components of the estimated streams to individual streams. Then, each agent extracts sound features from the residual input by using consistent attributes.

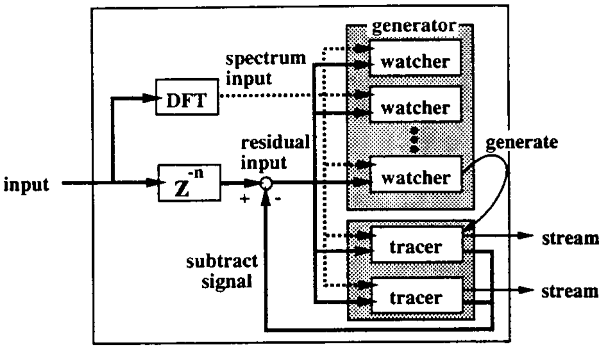

Figure 1: Watchers detect new streams and generate tracers, while tracers extract the streams

Instead of the auditory model, we used the Fourier transformation in our system because it is fully analyzed and easy to implement. We believe many important auditory features can be extracted by Fourier analysis.

### 3.2. Design of agents

We designed four agents to segregate harmonic sounds and a background noise. A pitch watcher detects a harmonic stream such as a voiced speech, and a harmonics tracer extracts it. The noise watcher detects a static noise, and the noise tracer segregates it. Harmonic sounds are segregated by means of a waveform while a static noise is segregated by means of average spectrum intensity. These agents have adaptive mechanisms to segregate more than one dynamic harmonic sound at the same time, although they have a simple noise manipulation mechanism that treats it as a single static noise.

We shall use several terms that should be defined. The first is the *harmonic intensity* $E_t(\omega)$ of the sound wave $x_t(\tau)$ at frame $t$. It is defined as

$$E_t(\omega) = \sum_k \| H_{t,k}(\omega) \|^2,$$

$$\text{where} \qquad H_{t,k}(\omega) = \sum_\tau x_t(\tau) \cdot \exp(-jk\omega\tau),$$

$\tau$ is time, $k$ is the index of harmonics, $x_t(\tau)$ is the residual input, and $H_{t,k}(\omega)$ is the sound component of the $k$th overtone. For the sound consistency check, we use a *valid overtone* for the harmonic stream and *valid period* for the noise stream. An overtone of a pitch watcher or of a harmonics tracer is *valid* if Equation 1 and Equation 2 hold. Valid overtones are used as segregation keys by each agent. The period is *valid* if there are no sound streams detected other than a background noise. The average spectrum intensity of the noise is calculated during the valid period. We also use *valid harmonic intensity*, $E'_t(\omega)$, which is defined as the sum of the $\| H_{t,k}(\omega) \|$ of valid overtones.

We define an overtone as *valid* if the intensity of the overtone is larger than a threshold (Equation 1), and if the local time transition of the intensity can be approximated in a linear fashion (Equation 2). Current setting of values is also shown.

$$\| H_{t,k}(\omega) \| > c \cdot \| \mathrm{DFT}_t(k \cdot \omega) \|, \ (c = 0.15), \qquad (1)$$

$$\sigma^2_{t,k} < p \cdot M_{t,k}, \ (p = 0.05). \qquad (2)$$

where $\mathrm{DFT}_t(\omega)$ is the frequency component of the spectrum input at frequency $\omega$, frame $t$, $\sigma^2_{t,k}$ and $M_{t,k}$ are the following calculated values: each harmonics tracer and each pitch watcher estimates the frequency $\bar{\omega}_{\tau,k}$ of the $k$th overtone at frame $\tau$ ($= t, \ldots, t+m-1$) using the spectrum input that contains $m$ ($=$ 10) input frames, when each residual input of frame $t$ arrives. For this estimate, each agent first extracts fundamental frequencies by tracking spectrum peaks as its overtones on the spectrum input. Thus, the sum of the peak intensities is maximized at each frame within the limited fundamental frequency region neighboring the fundamental frequency of the previous frame. If two or more agents take the same peak as their overtone, the agents retrack peaks excluding that peak, except for the tracer whose corresponding overtone (or peak) at the previous time frame has the most similar intensity to the intensity of that peak. Then, $\bar{\omega}_{\tau,k}$ is the frequency of the tracked peak for the $k$th overtone if there is a tracked peak for the $k$th overtone. Otherwise, $\bar{\omega}_{\tau,k}$ is estimated using the frequency of the other peaks. Next, $M_{t,k}$ is calculated as the mean value of $\| \bar{H}_{n,k}(\bar{\omega}_{n,k}) \|$, (n = t,...,t+m-1), and $\sigma^2_{t,k}$ is calculated as the variance of $\| \bar{H}_{n+1,k}(\bar{\omega}_{n+1,k}) - \bar{H}_{n,k}(\bar{\omega}_{n,k}) \|, (n = t, \ldots, t+m-2)$, where $\bar{H}_{n,k}(\bar{\omega}_{n,k}) = \| \mathrm{DFT}_n(k \cdot \bar{\omega}_{n,k}) \|$.

### 3.3. Pitch watcher

Many pitch watchers are included in the generator. Each pitch watcher has its own frequency region (about 25 Hz in width) and watches to see whether a new harmonic stream, whose fundamental frequency is in the region, appears at each residual input. This watcher is activated when the following conditions are satisfied: (a) $E'_t(\omega)/E_t(\omega) > r$ ($r = 0.1$), and (b) there is a power peak near frequency $\omega$ in the residual signal, where $\omega$ is the frequency that maximizes $E_t(\omega)$ within the region. Of all the active pitch watchers, the pitch watcher that gives the maximum $E_t(\omega)$ generates a new tracer.

### 3.4. Noise watcher

The noise watcher generates a noise tracer when it detects a noise signal. It generates the noise tracer when the power of the input signal becomes more than a threshold during the valid period and no pitch watcher is activated. In the current implementation, the number of generated noise tracers is at most one.

### 3.5. Harmonics tracer

A harmonics tracer gets the initial fundamental frequency from a pitch watcher when it is generated. At each residual input, each harmonics tracer extracts the fundamental frequency that maximizes the valid harmonic intensity, $E'_t(\omega)$. It then calculates the intensity and the phase of each overtone by evaluating the absolute value and the phase of $H_{t,k}(\omega)$. It creates subtract signals in a waveform by adjusting the phase of its overtones to the phase of the next input frame, and recovers its signal by adding its subtract signal to the residual signal before calculating the fundamental frequency. If there are no longer valid overtones, or if the intensity of the fundamental overtone drops below a threshold value, it terminates itself.

### 3.6. Noise tracer

The noise tracer segregates the static noise stream by means of the average spectrum intensity [9]. It calculates the spectrum intensity time average of the residual input during the valid period. The noise tracer sends subtract signals to other agents by means of the spectrum intensity. When a tracer receives a spectrum subtract signal, it estimates the intensities of its sound components at each frequency by subtracting the signal values. The subtract signal of the noise tracer inhibits the generator from generating unnecessary tracers and makes harmonics tracers robust against a nonharmonic noise. The noise tracer calculates average spectrum intensity for a long-time range as well as for a short-time range. It terminates itself when the short-time range average intensity drops below a threshold.

## 4. EVALUATION

We evaluated the system by using a mixture of a male voiced speech and a female voiced speech, both saying "aiueo" (Figure 2(a)). Segregated streams are shown

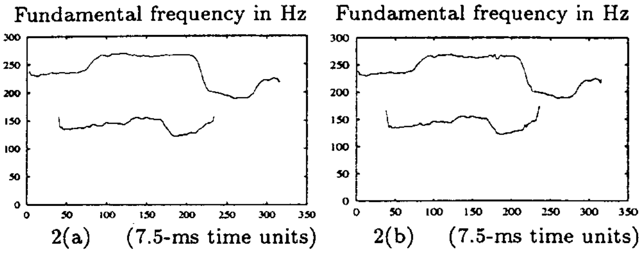

Figure 2: Fundamental frequency patterns of sound streams, (a) input mixture of male speech (lower curve) and female speech (upper one), (b) segregated male and female speech.

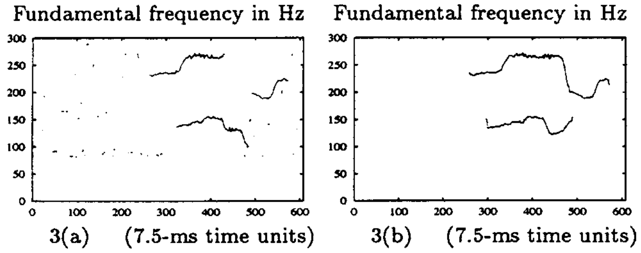

Figure 3: (a) Segregated streams under white nose whose power is the same as that of male speech without noise tracer and (b) with noise tracer. Noise tracer improves stream segregation performance.

in Figure 2(b). The quality of the segregated sounds is good. Under white noise, the performance of the system without the noise tracer and with it are shown in Figures 3(a) and 3(b). The total number of tracers with and without the noise tracer was 13 and 55. Using the old system, the total number was 46 and 268 with and without the noise tracer. The present system effectively reduced the number of tracers both with and without the noise tracer compared to the old system.

The quality of segregated sound streams was evaluated with regard to spectral distortion and pitch error. Results are shown in Figures 4(a) and 4(b). This evaluation proves that the noise tracer improves segregation quality even in a noisy environment.

## 5. CONCLUDING REMARKS

We presented a computational approach to sound stream segregation based on a multi-agent paradigm. The new system uses four kinds of agents, pitch watchers, a noise watcher, harmonic tracers, and a noise tracer. It can trace any number of streams in a sound mixture. The system can segregate a male and a female speech in a noisy environment. The results suggest a clue to understanding the *cocktail party problem*. We will next investigate this problem by designing a new agent that extracts only a human voice, including consonants, by using the information extracted by harmonics tracers.

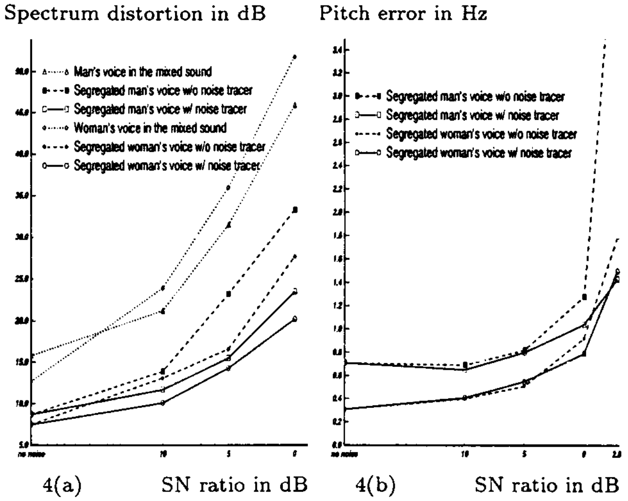

Figure 4: Quality of segregated sound streams under white noise, (a) spectrum distortion of segregated and mixed voices, (b) pitch error of segregated voices. SN ratio of white noise to male speech is 10 db, 5 db, 0 db and -2.8 db. Spectral error for every segregated sound is reduced by less than half from its original sound in sound mixture. Noise tracer reduces spectral error by 2.2 to 9.8 db. Pitch error without noise tracer is small while noise level is low, but error increases as noise level becomes higher.

## 6. REFERENCES

[1] Bregman: *Auditory Scene Analysis – the perceptual organization of sound*, MIT Press, '90.

[2] Brown: Computational auditory scene analysis: A representational approach, *PhD thesis*, Univ.of Sheffield, '92.

[3] Slaney, Naar, and Lyon: Auditory Model Inversion For Sound Separation. *Proc. of ICASSP-94*.

[4] Cooke, Brown, Crawford, and Green: Computational Auditory Scene Analysis: listening to several things at once. *Endeavour*, **17:4**, '93.

[5] Costas: Residual Signal Analysis–A Search and Destroy Approach to Spectral Analysis. *Proc. of 1st ASSP Workshop on Spectral Estimation*, '81.

[6] Ramalingam and Kumaresan: Voiced-speech analysis based on the residual interfering signal canceler (RISC) algorithm. *Proc. of ICASSP-94*.

[7] Warren, Obusek, and Ackroff: Auditory induction: Perceptual synthesis of absent sounds. *Science*, Vol.176, pp. 1149–1151 (1972).

[8] Nakatani, Okuno, and Kawabata: Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System. *Proc. of AAAI-94*.

[9] Boll: A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech, *ICASSP-79*.