Interfacing Sound Stream Segregation to Automatic Speech Recognition — Preliminary Results on Listening to Several Sounds Simultaneously

Hiroshi G. Okuno, Tomohiro Nakatani and **Takeshi Kawabata**

NTT Basic Research Laboratories
Nippon Telegraph and Telephone Corporation
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-01, JAPAN
okuno@nue.org nakatani@horn.brl.ntt.jp kaw@idea.brl.ntt.jp

Abstract

This paper reports the preliminary results of experiments on listening to several sounds at once. Two issues are addressed: segregating speech streams from a mixture of sounds, and interfacing speech stream segregation with automatic speech recognition (*ASR*). Speech stream segregation (*SSS*) is modeled as a process of extracting harmonic fragments, grouping these extracted harmonic fragments, and substituting some sounds for non-harmonic parts of groups. This system is implemented by extending the harmonic-based stream segregation system reported at AAAI-94 and IJCAI-95. The main problem in interfacing SSS with HMM-based ASR is how to improve recognition performance which is degraded by spectral distortion of segregated sounds caused mainly by the binaural input, grouping, and residue substitution. Our solution is to re-train the parameters of the HMM with training data binauralized for four directions, to group harmonic fragments according to their directions, and to substitute the residue of harmonic fragments for non-harmonic parts of each group. Experiments with 500 mixtures of two women's utterances of a word showed that the cumulative accuracy of word recognition up to the 10th candidate of each woman's utterance is, on average, 75%.

Introduction

Usually, people hear a mixture of sounds, and people with normal hearing can segregate sounds from the mixture and focus on a particular voice or sound in a noisy environment. This capability is known as the *cocktail party effect* (Cherry 1953). Perceptual segregation of sounds, called *auditory scene analysis*, has been studied by psychoacoustic researchers for more than forty years. Although many observations have been analyzed and reported (Bregman 1990), it is only recently that researchers have begun to use computer modeling of auditory scene analysis (Cooke et al. 1993; Green et al. 1995; Nakatani et al. 1994a). This emerging research area is called *computational auditory scene analysis (CASA)* and a workshop on CASA was held at IJCAI-95 (Rosenthal & Okuno 1996).

One application of CASA is as a front-end system for *automatic speech recognition* (*ASR*) systems. Hearing impaired people find it difficult to listen to sounds in a noisy environment. Sound segregation is expected to improve the performance of hearing aids by reducing background noises, echoes, and the sounds of competing talkers. Similarly, most current ASR systems do not work well in the presence of competing voices or interfering noises. CASA may provide a robust front-end for ASR systems.

CASA is not simply a hearing aid for ASR systems, though. Computer audition can listen to several things at once by segregating the sounds. This capability to listen to several things simultaneously has been called the *Prince Shotoku effect* by Okuno (Okuno et al. 1995) after Prince Shotoku (574–622 A.D.) who is said to have been able to listen to ten people's petitions at the same time. Since this is virtually impossible for humans to do, CASA research would make computer audition more powerful than human audition, similar to the relationship of an airplane's flying ability to that of a bird.

At present, one of the hottest topics of ASR research is how to make more robust ASR systems that perform well outside *laboratory conditions* (Hansen et al. 1994). Usually the approaches taken are to reduce noise, use speaker adaptation, and treat sounds other than human voices as noise. CASA takes an opposite approach. First, it deals with the problems of handling general sounds to develop methods and technologies. Then it applies these to develop ASR systems that work in a real world environment.

In this paper, we discuss the issues concerning interfacing of sound segregation systems with ASR systems and report preliminary results on ASR for a mixture of sounds.

Sound Stream Segregation

Sound segregation should be incremental, because CASA is used as a front-end system for ASR systems and other applications that should run in real-time. Many representations of a sound have been proposed, for example, auditory maps (Brown 1992) and synchrony strands (Cooke et al. 1993), but most of them are unsuitable for incremental processing. Nakatani and Okuno proposed using a *sound stream* (or simply *stream*) to represent a sound (Nakatani et al. 1994a). A sound stream is a group of sound components that have some consistent attributes. By using sound streams, the Prince Shotoku effect can be modeled as shown in Fig. 1. Sound streams are segregated by the sound segregation system, and then speech streams are selected and passed on to the ASR systems.

Sound stream segregation consists of two subprocesses:

1. **Stream fragment extraction** — a fragment of a stream that has the same consistent attributes is extracted from a mixture of sounds.

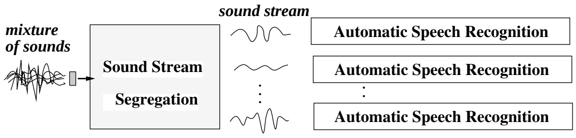

Figure 1: Modeling of the Prince Shotoku Effect or of Listening to several sounds simultaneously

2. **Stream fragment grouping** — stream fragments are grouped into a stream according to some consistent attributes.

Most sound segregation systems developed so far have limitations. Some systems assume the number of sounds, or the characteristics of sounds such as voice or music (e.g., (Ramalingam 1994)). Some run in a batch mode (e.g., (Brown 1992; Cooke et al. 1993)). Since CASA tries to manipulate any kind of sound, sound segregation should be able to manipulate any kind of sound. For that reason, sound segregation systems should work primarily with the low level characteristics of sound. Once the performance of such systems has been assessed, the use of higher level characteristics of sounds or combining bottom-up and top-down processing should be attempted.

Nakatani et al. used a harmonic structure [1] and the direction of the sound source as consistent attributes for segregation. They developed two systems: the harmonic-based stream segregation (*HBSS*) (Nakatani et al. 1994a; Nakatani et al. 1995a), and the binaural harmonic-based stream segregation (*Bi-HBSS*) systems (Nakatani et al. 1996). Both systems were designed and implemented in a multi-agent system with the residue-driven architecture (Nakatani et al. 1995b). We adopted these two systems to extract stream fragments from a mixture of sounds, since they run incrementally by using lower level sound characteristics. This section explains in detail how HBSS and Bi-HBSS work.

Harmonic-based Sound Segregation

The HBSS uses three kinds of agents: an event-detector, a tracer-generator, and tracers (Fig. 2) (Nakatani et al. 1994a; Nakatani et al. 1995a). It works as follows:

1. An event-detector subtracts a set of predicted inputs from the actual input and sends the residue to the tracer-generator and tracers.
2. If the residue exceeds a threshold value, the tracer-generator searches for a harmonic structure in the residue. If it finds a harmonic structure and its fundamental stream, it generates a tracer to trace the harmonic structure.
3. Each tracer extracts a harmonic stream fragment by tracing the fundamental frequency of the stream. It also composes a predicted next input by adjusting the segregated stream fragment with the next input and sends this prediction to the event-detector.

[1] A harmonic structure consists of a fundamental frequency and its integer multiples or *overtones*.

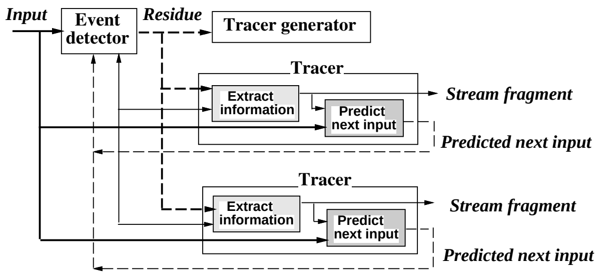

Figure 2: Harmonic-based Stream Segregation (HBSS)

Since tracers are dynamically generated and terminated in response to the input, a HBSS system can manipulate any number of sounds in principle. Of course, the setting of various thresholds determines the segregation performance.

The tracer-generator extracts a fundamental frequency from the residue of each time frame. For that purpose, the *harmonic intensity* $E_t(\omega)$ of the sound wave $x_t(\tau)$ at frame t is defined as

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

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

where τ is time, k is the index of harmonics, $x_t(\tau)$ is the residue, and $H_{t,k}(\omega)$ is the sound component of the kth overtone. Since some components of a harmonic structure are destroyed by other interfering sounds, not all overtones are reliable. Therefore, only a *valid overtone* for a harmonic structure is used. An overtone is defined as *valid* if the intensity of the overtone is larger than a threshold value and the time transition of the intensity can be locally approximated in a linear manner. The *valid harmonic intensity*, $E'_t(\omega)$, is also defined as the sum of the $\parallel H_{t,k}(\omega) \parallel$ of valid overtones.

When a (harmonic) tracer is generated, it gets the initial fundamental frequency from the tracer-generator, and at each time frame it extracts the fundamental frequency that maximizes the valid harmonic intensity $E_t(\omega)$. Then, it calculates the intensity and the phase of each overtone by evaluating the absolute value and that of $H_{t,k}(\omega)$ and extracts a stream fragment of the time frame. It also creates a predicted next input in a waveform by adjusting the phase of its overtones to the phase of the next input frame. If there are no longer valid overtones, or if the valid harmonic intensity drops below a threshold value, it terminates itself.

Binaural Harmonic-based Sound Segregation

When a mixture of sounds has harmonic structures whose fundamental frequencies are very close, HBSS may fail to segregate such sounds. For example, consider two harmonic sounds; one's fundamental frequency is increasing and the other's fundamental frequency is decreasing. When both fundamental frequencies cross, the HBSS cannot know whether two fundamental frequencies are crossing or approaching and departing. To cope with such problems and improve the segregation performance, binaural harmonic-based stream segregation (*Bi-HBSS*), which incorporates direction information into the HBSS, was proposed (Nakatani et al. 1996).

The Bi-HBSS takes a binaural input and extracts the direction of the sound source by calculating the interaural time difference (*ITD*) and interaural intensity difference (*IID*). More precisely, the Bi-HBSS uses two separate HBSS's for the right and left channels of the binaural input to extract harmonic stream fragments. Then, it calculates the ITD and IID by using a pair of harmonic stream fragments segregated. This method of calculating the ITD and IID reduces the computational costs, which is an important advantage since these values are usually calculated over the entire frequency region (Blauert 1983; Bodden 1993; Stadler & Rabinowitz 1993). The Bi-HBSS also utilizes the direction of the sound source to refine the harmonic structure by incorporating the direction into the validity. Thus, Bi-HBSS extracts a harmonic stream fragment and its direction. Internally, direction is represented by ITD (msec) and fundamental frequency is represented by *cent*. The unit, cent, is a logarithmic representation of frequency and 1 octave is equivalent to 1,200 cent.

The Bi-HBSS improves the segregation performance of the HBSS (Nakatani et al. 1995b; Nakatani et al. 1996). In addition, the spectral distortion of segregated sounds became very small when benchmarking was used with various mixtures of two women's utterances of Japanese vowels and interfering sounds (Nakatani et al. 1996).

However, the usage of binaural inputs may causes spectral distortion, because the spectrum of a binaural input is not the same as that of the original sound due to the shape of the human head. Such transformation is called the *head-related transfer function (HRTF)* (Blauert 1983). Due to the HRTF, the power of lower frequencies is usually decreased while that of higher frequencies is increased. Thus, it may make it difficulty to segregate a person's speech. The literature mentioned above did not examine this possibility.

Design of Speech Stream Segregation

Neither HBSS nor Bi-HBSS can segregate a speech stream, because it contains non-harmonic structures (e.g., consonants, especially unvoiced consonants) as well as harmonic structures (e.g., vowels and some voiced consonants). In this paper, we propose a simple method to extract a speech stream. First, the harmonic structures (vowels and some voiced consonants) of each stream are extracted by HBSS or Bi-HBSS and reconstructed by grouping. This process is called **harmonic grouping**. Second, non-harmonic structures (or most consonants) are reconstructed by substituting the residue. This process is called **residue substitution**. These processes also work incrementally, like the stream fragment extraction process. Note that in this scheme, consonants are extracted implicitly.

Harmonic Grouping Suppose that a new harmonic stream fragment ϕ is to be grouped. Let f_ϕ be the fundamental frequency of ϕ. The harmonic part of a stream is reconstructed in one of the following three ways (Nakatani et al. 1996; Rosenthal & Okuno 1996):

1. **F-grouping** — according to the nearness of the fundamental frequencies. Find an existing group, say Ψ, such that the difference $\mid f_\phi - f_\Psi \mid < \delta$. The value of δ is 300 cent if other new stream fragments exist at the same time with ϕ, 600 cent otherwise. If more than one existing groups are found, ϕ is grouped into the group that is the closest to f_phi. If only one existing group is found, ϕ is grouped into Ψ. Otherwise, ϕ forms a new group.
2. **D-grouping** — according to the nearness of the directions of the sound source. The range of nearness in ITD is 0.167 msec, which corresponds roughly to 20°. The algorithm is the same as the F-grouping.
3. **B-grouping** — If a stream fragment, ϕ, satisfies the above two conditions for a group, Ψ, it is grouped into Ψ. However, if ϕ has more than one such group, a group of minimum combined nearness is selected. The combined nearness, K, is defined as follows:

$$K = \alpha \frac{\mid \triangle f \mid}{c_f} + (1 - \alpha) \frac{\mid \triangle d \mid}{c_d}$$

where $c_f = 300$ cent, and $c_d = 0.167$ msec. The current value of the normalized factor, α, is 0.47.

The grouping is controlled by the gap threshold; if the time gap between two consecutive stream fragments is less than the gap threshold, they are grouped together with information about the missing components. The current value of the gap threshold is 500 msec, which is determined by the maximum duration of the consonants in the utterance database. Since HBSS extracts only harmonic structures, only F-grouping is applicable.

Residue substitution The idea behind the residue substitution is based on the observation that human listeners can perceptually restore a missing sound component if it is very brief and replaced by appropriate sounds. This auditory mechanism of phonemic restoration is known as *auditory induction* (Warren 1970). After harmonic grouping, harmonic components are included in a segregated stream, while non-harmonic components are left out. Since the missing components are non-harmonic, they cannot be extracted by either HBSS or Bi-HBSS and remain in the residue. Therefore, the missing components may be restored by substituting the residue produced by HBSS or Bi-HBSS.

The residue substitution, or which part of the residue is substituted for missing components, may be done by one of the following methods:

1. **All-residue substitution** — All the residue is used.
2. **Own-residue substitution** — Only the residue from the direction of the sound source is used.

In this paper, the former method is used, because the latter requires a precise determination of the sound source direction and thus the computational cost of separation is higher. In addition, the recognition performance of the latter is lower than that of the former, as will be shown later.

Issues in Interfacing SSS with ASR

We use a speech recognition system based on a hidden Markov model-based (*HMM*). An HMM usually uses the three characteristics in speech recognition; a spectral envelop, a pitch or a fundamental frequency, and a label or a pair consisting of the onset and offset times of speech. Since the input is a mixture of sounds, these characteristic, in particular the spectral envelop, are critically affected. Therefore, the recognition performance with a mixture of sounds is severely degraded by spectral distortion caused by interfering and competing sounds.

The segregation of the speech streams is intended to reduce the degradation, and is considered effective in recovering spectral distortion from a mixture of sounds. However, it also introduces another kind of spectral distortion to segregated streams, which is caused by extracting the harmonic structure, the head-related transfer function, or a binaural input, and the grouping and residue substitution. In the next section, the degradation of the recognition performance caused by segregation will be assessed and methods of recovery will be proposed.

The pitch error of Bi-HBSS for simple benchmarks is small (Nakatani et al. 1996). However, its evaluation with larger benchmarks is also needed. The onset of a segregated stream is detected only from the harmonic structures in HBSS. Since the beginning and end of speech are usually comprised of non-harmonic structures, the onset and offset times are extended by 40 msec for sounds segregated by HBSS. Since Bi-HBSS can detect whether a leading and/or trailing sound exists according to the directional information, the onset and offset is determined by this.

Influence of SSS on ASR

In this section, we assess the effect of segregation on ASR and propose methods to reduce this effect.

The ASR system used in this paper

The "HMM-LR" developed by ATR Inc. (Kita et al. 1990) is used system in this paper. The HMM-LR is a continuous speech recognition system that uses generalized LR parsing with a single discrete codebook. The size of the codebook is 256 and it was created from a set of standard data. The training and test data used in this paper were also created by ATR Inc. Since the primitive HMM-LR is a gender-dependent speech recognition system, HMM-LRs for male speakers (the *HMM-m*) and for female speakers (the *HMM-f*) were used. The parameters of each system were trained by using 5,240 words from five different sets of 1,048 utterances by each speaker. The recognition performance was evaluated by an open test, and 1,000 testing words were selected randomly from non-training data. The evaluation was based on word recognition. Therefore, the LR grammar for the HMM-m/f consists of only rules that the start symbol derives a terminal symbol directly. The evaluation measure used in this paper is the *cumulative accuracy up to the 10th candidate*, which specifies what percentage of words are recognized up to the 10th candidate by a particular HMM-LR. This measurement is popular for evaluating the actual speech recognition performance in a whole speech understanding system, because the top nth recognition candidates are used in successive language processors.

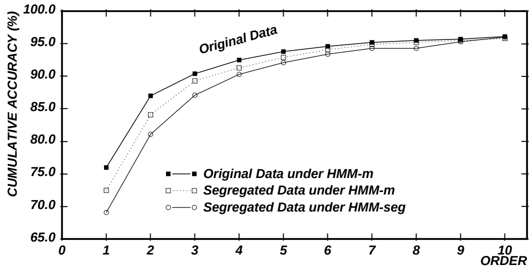

Figure 3: Influence of the harmonic structure extraction (Experiment 1)

Influence of the Harmonic Structure Extraction

To assess the influence of harmonic structure extraction on the word recognition performance, we have define a new operation called *harmonic structure reconstruction*, which is done as follows:

1. The HBSS extracts harmonic stream fragments from an utterance of a word by a single speaker.
2. All the extracted harmonic stream fragments are grouped into the same stream.
3. All the residue is substituted in the stream for the time frames where no harmonic structure was extracted.

Experiment 1: Harmonic structure reconstruction and word recognition was performed using the HMM-m for over 1,000 utterances of a word by a male speaker. The cumulative accuracy of the recognition is shown in Fig. 3. In Fig. 3, the curve denoted as the *original data* shows the recognition rate for the same original utterances by the same speaker. The word recognition rate was lower by 3.5% for the first candidate when the HMM-m was used, but was almost equal in cumulative accuracy for the 10th candidate. This demonstrates that the harmonic structure reconstruction has little effect on the word recognition performance.

We tried to improve the recognition rate by re-training the parameters of the HMM-LR by using all the training data provided through harmonic structure reconstruction. The resulting HMM-LR, called **HMM-seg**, however, did not improve the recognition rate as shown in Fig. 3. Therefore, we did not adopt any special treatment for harmonic structure reconstruction.

Influence of the Head-related Transfer Function

As we mentioned, a binaural sound is equivalent to its original sound transformed by a head-related transfer function (*HRTF*) with a particular direction.

Experiment 2: To evaluate the influence of the HRTF, all the test data were converted to binaural sounds as follows, and then recognized by the HMM-m.

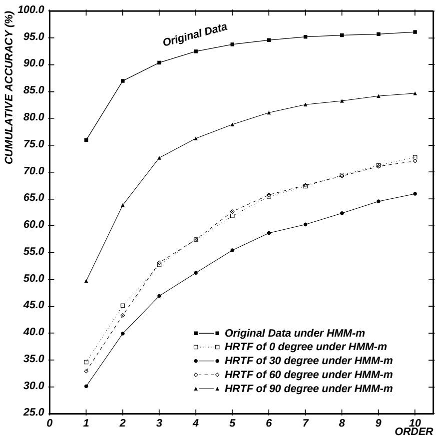

Figure 4: Influence of the head-related transfer function (Experiment 2)

1. HRTFs in four directions (0°, 30°, 60°, and 90°) [2] were applied to each test utterance to generate a binaural sound.
2. For each binaural sound, the monaural sound was extracted from the channel with the greater power, in this case, the left channel.
3. The power level was adjusted so that its average power was equivalent to that of the original sound. This operation is called *power adjustment*.
4. The resulting monaural sounds (the *HRTF'ed* test data) were given to the HMM-m for word recognition.

The cumulative recognition accuracy for the HRTF'ed test data is shown in Fig. 4. The original data is also shown for comparison. The decrease in the cumulative accuracy for the 10th candidate ranged from 11.4% to 30.1%. The size of the decrease depended on the direction of the sound source and was greatest for 30° and least for 90°.

Recovering the Performance Degradation caused by the HRTF

Two methods to recover the decrease in recognition accuracy caused by HRTF have been tried:

1. Re-training the HMM-LR parameters with the HTRF'ed training data, and
2. Correcting the frequency characteristics of the HRTF.

Re-training of the parameters of the HMM-LR We converted the training data for the HMM-LR parameters by applying the HRTF in the four directions to the training data with power adjustment. We refer to the re-trained HMM-LR for male speakers as the *HMM-hrtf-m*.

The cumulative recognition accuracy of the HRTF'ed test data by the HMM-hrtf-m is shown in Fig. 5. The decrease in the cumulative accuracy was significantly reduced and

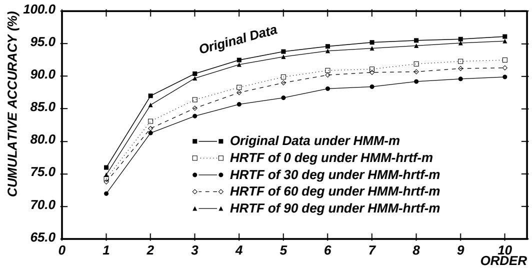

Figure 5: Recovery by re-trained HMM-LR

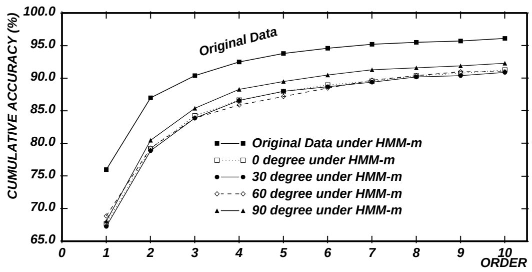

Figure 6: Recovery by correcting the F-char of HRTF

almost vanishes for 90°. However, the degradation still depended on the direction of the sound source.

Frequency Characteristics (F-Char) Correction The effect of the HRTF is to amplify the higher frequency region while attenuating the lower frequency region. For example, the Japanese word "aji" (taste) sounds like "ashi" (foot) if an HRTF of any degree is applied. To recover the spectral distortion caused by the HRTF, we corrected the frequency characteristics (F-Char) of the HRTR'ed test data through power adjustment. After this correction, the test data could be recognized by the HMM-m (Fig. 6). The variance in the recognition rate due to different directions was resolved, but the overall improvement was not as great as with the HMM-hrtf-m.

Since the latter method requires a precise determination of the directions, though, it cannot be used when the sound source is moving. In addition, the amount of HRTF data for the various directions is very large and its computational cost is significant. Therefore, we used the HMM-hrtf-m/f to recognize binaural data.

Influence of the Harmonic Grouping and Residue Substitution

Experiment 3: The influence of harmonic grouping by the F-grouping, D-grouping, and B-grouping was evaluated by the following method:

1. The Bi-HBSS extracted harmonic stream fragments from binaural input in four directions (0°, 30°, 60°, and 90°) for a man's utterance.
2. Sound stream fragments were grouped into a stream by one of the three groupings and the non-harmonic components of the stream are filled in through the all-residue substitution.

[2] The angle is calculated counterclockwise from the center, and thus 0°, 90°, and -90° mean the center, the leftmost and the rightmost, respectively.

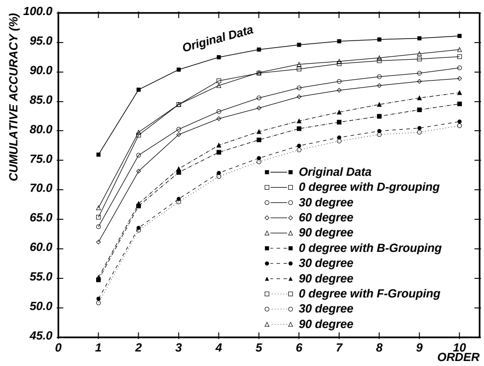

Figure 7: Influence of harmonic grouping (Experiment 3)

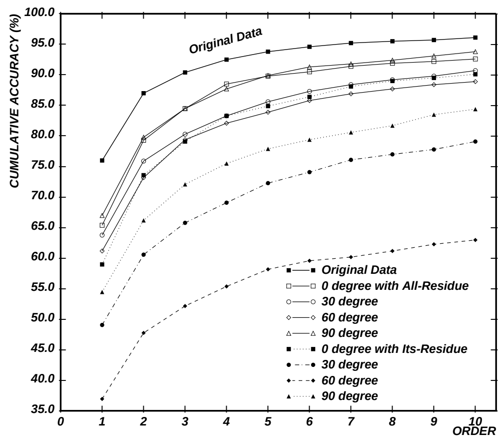

Figure 8: Influence of residue substitution (Experiment 4)

3. Power adjustment was applied to the segregated sound streams.
4. The resulting sounds were recognized with the HMM-hrtf-m.

The recognition rate is shown in Fig. 7. The best performance was with the D-grouping, while the worst was with the F-grouping. The recognition with the F-grouping was poor because only the previous state of the fundamental frequency was used to group stream fragments. This also led to poor performance with the B-grouping. Longer temporal characteristics of a fundamental frequency should be used, but this remains for future work. Therefore, we adopted the D-grouping for the experiments described in the remainder of this paper.

Experiment 4: We evaluated the effect of residue substitution by either all-residue substitution or own-residue substitution in the same way as Experiment 3. The resulting recognition rates are shown in Fig. 8. The recognition rate was higher with the all-residue substitution than with the own-residue substitution. This is partially because the signals substituted by the own-residue were weaker than those by the all-residue. Therefore, we will use the all-residue substitution throughout the remainder of this paper.

Experiments on Listening to a Sound Mixture

Our assessment of the effect of segregation on ASR suggests that we should use Bi-HBSS with the D-grouping and the all-residue substitution and that segregated speech streams should be recognized by the HMM-hrtf-m/f. We also evaluated monaural segregation by HBSS with the all-residue substitution and the HMM-m/f. The experiments on recognizing a mixture of sounds were done under the following conditions:

1. The first speaker is 30° to the left of the center and utters a word first.
2. The second speaker is 30° to the right of the center and utters a word 150 msec after the first speaker.
3. There were 500 two-word testing combinations.
4. The power adjustment was not applied to any segregated sound, because the system cannot determine the original sound that corresponds to a segregated sound.

The utterance of the second speaker was delayed by 150 msec because the mixture of sounds was to be recognized directly by the HMM-m/f. Note that the actual first utterance is not always by the first speaker, although that is usually the case.

Listening to Two Sounds at the Same Time

Since the HMM-LR framework we used is gender-dependent, the following three benchmarks were used (see Table 1). The cumulative accuracies of recognition of the original data for Woman 1, Woman 2, Man 1, and Man 2 by the HMM-m/f were 94.19%, 95.10%, 94.99%, and 96.10%, respectively.

The recognition rate was measured without segregation, with segregation by HBSS, and with segregation by Bi-HBSS. The recognition performance in terms of cumulative accuracy up to the 10th candidate is summarized in Tables 2 to 4. The recognition performance of speech segregated by Bi-HBSS was better than when HBSS was used. With Bi-HBSS, the decrease in the recognition rate of the second woman's utterance from that of the original sound was 21.20%. Since these utterances could not be recognized at all without segregation, the error rate was reduced by 73.60% on average by the segregation.

Without segregation, the utterances of the first speaker could be recognized up to 37% if the label (the onset and offset times) was given by some means. In this experiment, the original labels created by human listeners at ATR were used. However, the recognition rate falls to almost zero when another sound is interfering (see the following experiments in Table 6 and 7).

The Bi-HBSS reduces the recognition errors of HBSS by 48.1%, 22.7%, and 23.1% for benchmarks 1, 2, and 3, respectively. The improvement for benchmark 1 is especially large because the frequency region of women's utterances is so narrow that their recognition is prone to recognition errors. Men's utterances, in particular, the second man's utterances of benchmark 3, are not well segregated by HBSS or Bi-HBSS. The pitch of the second

Table 1: Benchmark sounds 1–3

Benchmark No.	Speaker 1	Speaker 2
1	Woman 1	Woman 2
2	Man 1	Woman 2
3	Man 1	Man 2

Table 2: Recognition Rate of Benchmark 1

	10th Cumulative Accuracy		
	Average	Speaker 1	Speaker 2
No segregation	—	18.80%	0.60%
HBSS	27.51%	26.31%	28.71%
Bi-HBSS	75.60%	77.31%	73.90%

Table 3: Recognition Rate of Benchmark 2

	10th Cumulative Accuracy		
	average	Speaker 1	Speaker 2
No segregation	—	36.40%	0.40%
HBSS	43.88%	47.19%	40.56%
Bi-HBSS	66.60%	62.25%	71.69%

Table 4: Recognition Rate of Benchmark 3

	10th Cumulative Accuracy		
	average	Speaker 1	Speaker 2
No segregation	—	37.80%	0.40%
HBSS	28.11%	31.73%	24.50%
Bi-HBSS	53.21%	61.24%	45.18%

man is less than 100 Hz while that of the first man is about 110 Hz. Lower pitch is more difficult to segregate.

Listening to Three Sounds at the Same Time

Our next experiment was to segregate speech streams from a mixture of three sounds. Two benchmarks were composed by adding an intermittent sound to the sounds of benchmark 1 (see Table 5). The intermittent sound was a harmonic sound with a 250 Hz fundamental frequency that was repeated for 1,000 msec at 50 msec intervals. Its direction was 0°, that is, from the center. The signal-to-noise ratio (SNR) of the woman's utterance to the intermittent sound was 1.7 dB and -1.3 dB, respectively, for benchmark 4 and 5. The actual SNR was further reduced, because the other woman's utterance was also an interfering sound.

The recognition performance in terms of 10th cumulative accuracy are summarized in Tables 6 and 7. The degradation with HBSS and Bi-HBSS caused by the intermittent sound of benchmark 4 was 7.9% and 23.3%, respectively. When the power of the intermittent sound was amplified and the SNR of the woman's utterances decreased by 3 dB, the additional degradation with HBSS and Bi-HBSS was 1.5% and 5.8%, respectively. Segregation by either HBSS or Bi-HBSS seems robust against an increase in the power level of interfering sounds.

Discussion and Future work

In this paper, we have described our experiments on the Prince Shotoku effect, or listening to several sounds simultaneously. We would like to make the following observations.

Table 5: Benchmark sounds 4–5

No.	1st Speaker (SNR)	2nd Speaker (SNR)	3rd Sound
4	Woman 1 (1.7 dB)	Woman 2 (1.7 dB)	Intermittent Sound
5	Woman 1 (-1.3 dB)	Woman 2 (-1.3 dB)	Intermittent Sound

Table 6: Recognition Rate of Benchmark 4

	10th Cumulative Accuracy		
	average	Speaker 1	Speaker 2
No segregation	—	0.00%	0.40%
HBSS	19.58%	16.87%	22.29%
Bi-HBSS	52.31%	57.63%	46.99%

Table 7: Recognition Rate of Benchmark 5

	10th Cumulative Accuracy		
	average	Speaker 1	Speaker 2
No segregation	—	0.00%	0.20%
HBSS	18.07%	14.26%	21.89%
Bi-HBSS	46.48%	49.09%	43.86%

(1) Most of the sound stream segregation systems developed so far (Bodden 1993; Brown 1992; Cooke et al. 1993; Green et al. 1995; Ramalingam 1994) run in batch. However, HBSS and Bi-HBSS systems run incrementally, which is expected to make them easier to run in real time.

(2) Directional information can be extracted by binaural input (Blauert 1983; Bodden 1993) or by microphone arrays (Hansen et al. 1994; Stadler & Rabinowitz 1993). Our results prove the effectiveness of localization by using a binaural input. However, this severely degrades the recognition rate due to spectral distortion; this has not been reported in the literature as far as we know. Therefore, we are currently engaged in designing a sophisticated mechanism to integrate HBSS and Bi-HBSS to overcome the drawbacks caused by a binaural input.

(3) The method to extract a speech with consonants is based on auditory induction, a psychacoustical observation. This method is considered as the first approximation for speech stream segregation, because it does not use any characteristics specific to human voices, e.g., formants. In addition, we should attempt to incorporate a wider set of the segregation and grouping phenomena of psychoacoustics such as common onset, offset, AM and FM modulations, formants, and localization such as elevation and azimuth.

(4) In HMM-based speech recognition systems, the leading part of a sound is very important to focus the search and if the leading part is missing, the recognition fails. Examination of the recognition patterns shows that the latter part of a word or a component of a complex word is often clearly recognized, but this is still treated as failure.

(5) Since a fragment of a word is more accurately segregated than the whole word, top-down processing is expected to play an important role in the recognition. Various methods developed for speech understanding systems should be

used to improve the recognition and understanding.

(6) In this paper, we used standard discrete-type hidden Markov models for an initial assessment. However, HMM technologies have been improved in recent years, especially in terms of their robustness (Hansen et al. 1994; Minami & Furui 1995). The evaluation of our SSS in sophisticated HMM frameworks remains as future work.

(7) Our approach is bottom-up, primarily because one goal of our research is to identify the capability and limitations of the bottom-up approach. However, the top-down approach is also needed for CASA, because a human listener's knowledge and experience plays an essential role in listening and understanding (Handel 1989).

(8) To integrate bottom-up and top-down processes, system architecture is essential. The HBSS and Bi-HBSS systems are modeled on the residue-driven architecture with multi-agent systems. These systems can be extended for such integration by using subsumption architecture (Nakatani et al. 1994a). A common system architecture for such integration is the black board architecture (Cooke et al. 1993; Lesser et al. 1993). The modeling of CASA represents an important area for future work.

Conclusions

This paper reported the preliminary results of experiments on listening to several sounds at once. We proposed the segregation of speech streams by extracting and grouping harmonic stream fragments while substituting the residue for non-harmonic components. Since the segregation system uses a binaural input, it can interface with the hidden Markov model-based speech recognition systems only if the training data is converted to binaural data.

Experiments with 500 mixtures of two women's utterances of a word showed that the 10th cumulative accuracy of speech recognition of each woman's utterance is, on average, 75%. This performance was attained without using any features specific to human voices. Therefore, this result should encourage the AI community to engage more actively in computational auditory scene analysis (CASA) and computer audition. In addition, because audition is more dependent on the listener's knowledge and experience than vision, we believe that more attention should be paid to CASA.

Acknowledgments

We thank Kunio Kashino, Masataka Goto, Norihiro Hagita and Ken'ichiro Ishii for their valuable discussions.

References

Blauert, J. 1983. *Spatial Hearing: the Psychophysics of Human Sound Localization*. MIT Press.

Bodden, M. 1993. Modeling human sound-source localization and the cocktail-party-effect. *Acta Acustica* 1:43–55.

Bregman, A.S. 1990. *Auditory Scene Analysis – the Perceptual Organization of Sound*. MIT Press.

Brown, G.J. 1992. Computational auditory scene analysis: A representational approach. Ph.D diss., Dept. of Computer Science, University of Sheffield.

Cherry, E.C. 1953. Some experiments on the recognition of speech, with one and with two ears. *Journal of Acoustic Society of America* 25:975–979.

Cooke, M.P.; Brown, G.J.; Crawford, M.; and Green, P. 1993. Computational Auditory Scene Analysis: listening to several things at once. *Endeavour*, 17(4):186–190.

Handel, S. 1989. *Listening – An Introduction to the Perception of Auditory Events*. MIT Press.

Hansen, J.H.L.; Mammone, R.J.; and Young, S. 1994. Editorial for the special issue of the IEEE transactions on speech and audio processing on robust speech processing". *Transactions on Speech and Audio Processing* 2(4):549-550.

Green, P.D.; Cooke, M.P.; and Crawford, M.D. 1995. Auditory Scene Analysis and Hidden Markov Model Recognition of Speech in Noise. In Proceedings of 1995 International Conference on Acoustics, Speech and Signal Processing, Vol.1:401–404, IEEE.

Kita, K.; Kawabata, T.; and Shikano, H. 1990. HMM continuous speech recognition using generalized LR parsing. Transactions of the Information Processing Society of Japan 31(3):472–480.

Lesser, V.; Nawab, S.H.; Gallastegi, I.; and Klassner, F. 1993. IPUS: An Architecture for Integrated Signal Processing and Signal Interpretation in Complex Environments. In Proceedings of the Eleventh National Conference on Artificial Intelligence, 249–255.

Minami, Y, and Furui, S. 1995. A Maximum Likelihood Procedure for A Universal Adaptation Method based on HMM Composition. In Proceedings of 1995 International Conference on Acoustics, Speech and Signal Processing, Vol.1:129–132, IEEE.

Nakatani, T.; Okuno, H.G.; and Kawabata, T. 1994a. Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System. In Proceedings of the Twelfth National Conference on Artificial Intelligence, 100–107, AAAI.

Nakatani, T.; Kawabata, T.; and Okuno, H.G. 1995a. A computational model of sound stream segregation with the multi-agent paradigm. In Proceedings of 1995 International Conference on Acoustics, Speech and Signal Processing, Vol.4:2671–2674, IEEE.

Nakatani, T.; Okuno, H.G.; and Kawabata, T. 1995b. Residue-driven architecture for Computational Auditory Scene Analysis. In Proceedings of the th International Joint Conference on Artificial Intelligence, Vol.1:165–172, IJCAI.

Nakatani, T.; Goto, M.; and Okuno, H.G. 1996. Localization by harmonic structure and its application to harmonic sound stream segregation. In Proceedings of 1996 International Conference on Acoustics, Speech and Signal Processing, IEEE. Forthcoming.

Okuno, H.G.; Nakatani, T.; and Kawabata, T. 1995. Cocktail-Party Effect with Computational Auditory Scene Analysis — Preliminary Report —. In *Symbiosis of Human and Artifact — Proceedings of the Sixth International Conference on Human-Computer Interaction*, Vol.2:503–508, Elsevier Science B.V.

Ramalingam, C.S., and Kumaresan, R. 1994. Voiced-speech analysis based on the residual interfering signal canceler (RISC) algorithm. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Vol.I:473–476, IEEE.

Rosenthal, D., and Okuno, H.G. eidtors 1996. *Computational Auditory Scene Analysis*, LEA. Forthcoming.

Stadler, R.W., and Rabinowitz, W.M. 1993. On the potential of fixed arrays for hearing aids. *Journal of Acoustic Society of America* 94(3) Pt.1:1332–1342.

Warren, R.M. 1970. Perceptual restoration of missing speech sounds. *Science*167:392–393.