A NEW SPEECH ENHANCEMENT : SPEECH STREAM SEGREGATION

Hiroshi G. Okuno, Tomohiro Nakatani, and Takeshi Kawabata

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

ABSTRACT

Speech stream segregation is presented as a new speech enhancement for automatic speech recognition. Two issues are addressed: speech stream segregation from a mixture of sounds, and interfacing speech stream segregation with automatic speech recognition. Speech stream segregation is modeled as a process of extracting harmonic fragments, grouping these extracted harmonic fragments, and substituting non-harmonic residue for non-harmonic parts of groups. The main problem in interfacing speech stream segregation with HMM-based speech recognition is how to improve the degradation of recognition performance due to spectral distortion of segregated sounds, which is caused mainly by transfer function of a binaural input. Our solution is to re-train the parameters of HMM with training data binauralized for four directions. 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%.

1. INTRODUCTION

Sound is gathering attention as important media for multi-medal communications, but is less utilized as input media than characters or images. Since Cherry discovered the cocktail party effect in 1953, lots of observations on perceptual segregation of sound have been obtained in psychoacoustics and psychophysics. This research area is called *Auditory Scene Analysis* (ASA) [2]. However, the computer modeling of auditory scene analysis, or *Computational Auditory Scene Analysis* (CASA), has just started [3, 5, 18]. One application of CASA is a front-end for automatic speech recognition in real-world environments [16]. Automatic speech recognition systems developed to date attain high recognition performance in the laboratory environments, but may not work well in real-world environments.

To enable automatic speech recognition to work in such environments, *speech enhancement* is essential. Conventional approaches to speech enhancement are classified as noise reduction, speaker adaptation, or other robustness techniques [6, 9]. We take a novel approach to speech enhancement by applying CASA, that is, speech stream segregation. Speech stream segregation works as the front-end system for automatic speech recognition just as hearing aids for hearing impaired people.

Sound stream segregation is also considered as a new approach to hearing aids for human, because it is expected to enhance a sound, not restricted to a human voice, by reducing background noises, echoes and the sounds of competing talkers, and thus improve the performance of hearing aids.

Speech stream segregation is not simply a hearing aid for automatic speech recognition, though. Machine audition can listen to several things simultaneously by segregating each sound from a mixture of sounds as is shown in Fig. 1 [15, 16].

In this paper, we describe a speech stream segregation and present the interface between speech stream segregation and automatic speech recognition. Then, we report preliminary results on listening to a few things simultaneously.

2. DESIGN OF SPEECH STREAM SEGREGATION

Human voice consists of harmonic sounds such as vowel and voiced consonants, and non-harmonic sounds such as unvoiced consonants. By taking the structure of "Vowel (V) + Consonant (C) + Vowel (V)" of speech into consideration, we design the speech stream segregation as the following three subprocesses:

1. extracting harmonic stream fragments,
2. grouping harmonic stream fragments (*harmonic grouping*),
3. restoring non-harmonic parts by residue (*residue substitution*).

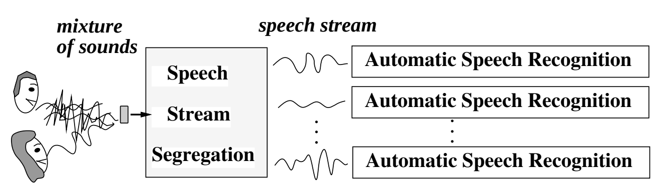

Figure 1: Modeling of listening to several speeches simultaneously. Speech streams are segregated from a mixture of sounds. Then, segregated speech streams are given to automatic speech recognition systems. If focus of attention mechanism is incorporated, cocktail party effect computer may be implemented by using this scheme.

2.1. Extracting Harmonic Stream Fragments

We use the Bi-HBSS (Binaural Harmonics-Based Stream Segregation) system [13, 14] to extract harmonic stream fragments from a binaural input. Bi-HBSS uses a harmonic structure and the direction of sound source as cues of segregation. Bi-HBSS adopts a pair of HBSSes [10, 11, 12] for the right and left channel to extract harmonic stream fragments. It determines the fundamental frequency (F_0) of a harmonic stream fragment by coordinating the pair of HBSSes. The direction of sound source is identified by calculating the interaural time difference (ITD) and interaural intensity difference (IID) of a pair of harmonic stream fragments of the same F_0 extracted by the pair of HBSS. Internally, the direction of sound source is represented by means of ITD. At the end of extracting a harmonic stream fragment, Bi-HBSS checks whether the fragment is followed by non-harmonic sounds of the same direction and sets the continuation flag if exists some.

Both Bi-HBSS and HBSS are designed by the residue-driven architecture [13], which subtracts harmonic structures from an input sound and uses the residue to check whether a new sound appears and to allocate a harmonic structure included in an input sound to an appropriate harmonic stream fragment exclusively [17]. Therefore, the residue include few harmonic structures in principle.

2.2. Harmonic Grouping

Since harmonic structure fragments carry information on the fundamental frequency and the direction of sound source, three criteria for grouping can be adopted; the proximity of fundamental frequency, the proximity of the direction of sound source, or the combination of these two proximity criteria. In this paper, we use the proximity of the directions of sound source, since it shows the best recognition performance [16]. The criteria of proximity of ITD is within 0.167 msec, which is roughly equivalent to 20°. If there are several groups that satisfy the criteria for a harmonic stream fragment, it is added to the closest group. If the continuation flag of the last stream fragment is set, a stream fragment that follows within the maximum time gap and satisfies the proximity criteria will be added to the same group. The maximum time gap is set to 150 msec which is the longest duration of a consonant in the database used in this paper. If there is no such group according to the proximity criteria, a new group is generated. The grouping mechanism is also designed by the residue-driven architecture [13, 14].

2.3. Residue Substitution

The residue obtained by subtracting harmonic structures from an input sound is substituted for non-harmonic parts of a group. If a group ends with non-harmonic parts, the residue is substituted for 150 msec. The idea of residue substitution is based on the psychophysical observation known as *auditory induction* [7, 20]. It is a phenomena that human listeners can perceptually restore a missing sound component if it is very brief and masked by appropriate sounds. There are several strict conditions for auditory induction to occur to human listeners, but we only use the condition that the masking sound is stronger that the masked sound, since our aim is to make speech recognition system to hear non-harmonic sounds.

3. INTERFACING SPEECH STREAM SEGREGATION WITH AUTOMATIC SPEECH RECOGNITION

The automatic speech recognition system used in this paper is HMM-LR developed at ATR [8]. HMM-LR uses a discrete single codebook whose size is 256. The codebook is generated by a set of standard phoneme data developed by ATR. Since it is gender-dependent, the parameters of HMM-LR for each gender is trained by a set of 5,240 words uttered by five speakers (each speaker utters different 1,000 words and common 240 words). Since *word recognition* is investigated in this paper, every rule for LR grammar generates a terminal symbol, i.e., a word, immediately from the start symbol.

We use the *cumulative accuracy up to the 10th candidate* (or simply *cumulative accuracy* in this paper) as the measure of the performance of word recognition. This is because a set of candidates in recognition process are usually given to successive speech understanding systems.

Automatic speech recognition by HMM in general uses *spectral envelop*, *pitch*, and *period* (a pair of onset and offset of a sound) as cues of recognition. (HMM-LR does not use information on pitch.) The distortion of any one of these cues may affect the performance of recognition. Within a mixture of sounds, these cues of any sound are distorted by the other sounds. Therefore, sound stream segregation removes such distortion well [12].

Speech stream segregation, however, introduces another kind of distortion. The spectrum of speech streams segregated by the proposed system is distorted severely by comparing their original single sound. This is caused by harmonic structure extraction, head-related transfer function (HRTF) of a binaural input, harmonic grouping and residue substitution. An open test of word recognition with 1,000 utterances of a single word proved that the degradation of recognition performance is the largest (20% to 40% depending on the direction of sound source) by HRTF and negligibly small (less than 10%) by the other causes [16].

To recover from the degradation of the performance of recognition caused by HRTF, two methods are investigated; re-training HMM parameters and correcting frequency-characteristics of segregated streams. The former reduces the degradation of the performance of recognition up to 3%, but the recognition performance still depends on the direction of sound source. The latter resolves the effects of the direction of sound source, but its improvement is not so large as by the former [16]. In addition, the latter requires the precise identification of the direction of sound source.

Therefore, we adopt re-training HMM parameters to interface the speech stream segregation with HMM-LR. First, a set of 5,240 utterances of a single word for training is binauralized analytically by using HTRF for four directions (0°, 30°, 60°, and 90°). Then, the parameters of each gender-dependent HMM-LR are re-trained with the binauralized training data. Since re-training is performed over 20,960 utterances, the robustness of HMM-LR against spectral distortion is increased.

4. EXPERIMENTS

4.1. Benchmarks for open test

We used five sets of 500 benchmark sounds; three sets for a mixture of two sounds and two sets for a mixture of three sounds (Table 1). The combinations of 500 mixture of two utterances are common. The first sound is uttered by the first speaker at 30° to the left from the center, and the second sound is uttered after 150 msec by the second speaker at 30° to the right from the center. To give a mixed sound directly to HMM-LR, the utterance of the second speaker is delayed by 150 msec. The original cumulative accuracy of word recognition uttered by a single speaker, Man 1, Man 2, Woman 1, and Woman 2, is 94.19%, 95.10%, 94.99%, and 96.10%, respectively.

Table 1: Five Sets of Benchmark Sounds

No	1st Sound (-30°)	2nd Sound (30°)	3rd Sound (0°)
1	Woman 1	Woman 2	—
2	Man 1	Woman 2	—
3	Man 1	Man 2	—
4	Woman 1	Woman 2	Intermittent
5	Woman 1	Woman 2	Intermittent

The third sound is an intermittent harmonic sounds from the center, whose F_0 is 250 Hz. It starts before the first speaker and repeats to last for 1 sec with 50 msec of pause. Benchmarks 4 and 5 are benchmarks 1 and 2 added by the third sound. The average power ratio of the first and second sounds to the third sound in benchmarks 4 and 5 is 1.7dB and -1.3dB, respectively. In addition, the total power of each mixture is reduced up to 4dB so that the maximum power can be represented by a 16-bit integer.

4.2. Listening to Two Sounds Simultaneously

The cumulative accuracy of word recognition up to the 10th candidate for benchmarks 1 to 3 is shown in Fig. 2–4. "**Proposed**" indicates results by the proposed method, "**Without Segregation**" indicates the results by recognizing the mixture of sounds without segregation, and "**Without Direction**" indicates the results by recognition of speech streams that are segregated by HBSS from a monaural input, and grouped by the proximity of fundamental frequency with residue substitution.

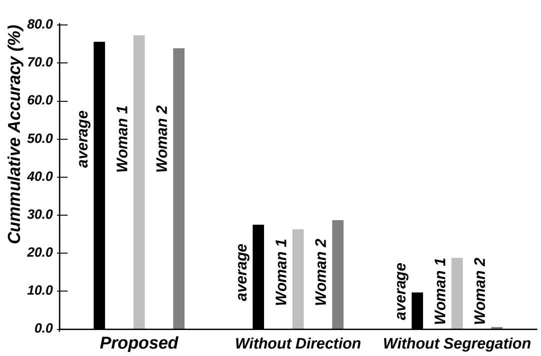

Figure 2: Cumulative Accuracy of recognition for Benchmark 1

The cumulative accuracy of word recognition for two women's utterances is 75.60% (Fig. 2). The cumulative accuracy for men's utterances is inferior to that for women's utterances, mainly because the lower F_0 makes segregation of harmonic structures more difficult. Since F_0 of Man 1's voice is about 100 Hz and its recognition with HBSS is poorer than that of without segregation (Fig. 4). However, the performance of its recognition with Bi-HBSS is improved by about 30%. These results shows that the leading part of segregated speech is essential for HMM recognition but the ending parts are not so because HMM may recover such errors.

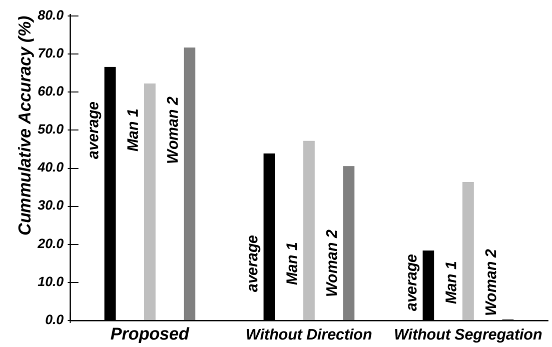

Figure 3: Cumulative Accuracy of recognition for Benchmark 2

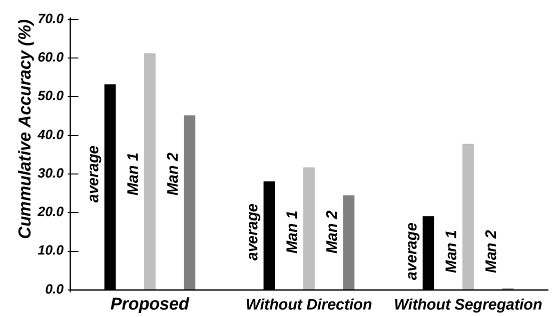

Figure 4: Cumulative Accuracy of recognition for Benchmark 3

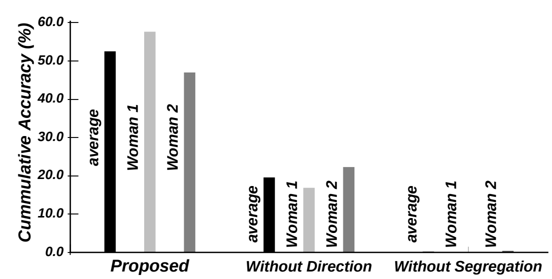

Figure 5: Cumulative Accuracy of recognition for Benchmark 4

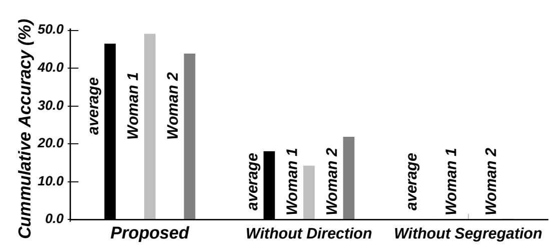

Figure 6: Cumulative Accuracy of recognition for Benchmark 5

4.3. Listening to 3 Sounds Simultaneously

The cumulative accuracy of word recognition up to the 10th candidate for benchmarks 4 and 5 is shown in Fig. 5 and 6, respectively. By the interfering intermittent sound, the average cumulative accuracy of word recognition is degraded by 23.30% (Figs.2 and 5). The signal noise ratio (SNR) of the first and second sounds to the third sound is 1.73dB, but from a viewpoint of one speaker, the other speaker's utterance is also an interfering sound and thus the actual SNR is much lower than 1.73dB. Although the SNR decreases further in benchmark 5, the average cumulative accuracy is not so degraded. Please note that without segregation did not work at all.

5. DISCUSSION AND FUTURE WORK

Although the performance of recognition reported in this paper is far from fulfilling requirements of real-world applications, it shows that only low level features of sounds such as a harmonic structure and the direction of sound source are effective in speech enhancement, and thus it encourages further research. Future work includes (1) using a stereo or multi-microphone input [19] instead of a binaural input [1] to identify the direction of sound source, (2) using voice-specific features such as tendency of F_0, formants, or speaker's model [9] to segregate speech streams, (3) integrating bottom-up and top-down processing, and (4) using recent automatic speech recognition systems.

6. CONCLUDING REMARKS

In this paper, speech stream segregation is presented as a new method for speech enhancement. Speech streams are segregation by extracting harmonic structures and by substituting non-harmonic residue of an input for non-harmonic parts. Since the proposed speech stream segregation uses a binaural input, the spectrum distortion of a segregated speech stream degrade the word recognition performance severely. By re-taining the HMM-LR parameters with binauralized data, the cumulative word recognition accuracy up to the 10-th candidates with 500 mixture sounds of two women's voices is improved by up to 72.3%

ACKNOWLEDGMENTS

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

7. REFERENCES

1. Bodden, M. "Modeling human sound-source localization and the cocktail-party-effect", *Acta Acustica* 1:43–55, 1993.
2. Bregman, A.S. *Auditory Scene Analysis – the Perceptual Organization of Sound.* MIT Press, 1990.
3. Brown, G.J. "Computational auditory scene analysis: A representational approach", *Ph.D diss.*, Dept. of Computer Science, University of Sheffield, 1992.
4. Cherry, E.C. "Some experiments on the recognition of speech, with one and with two ears", *J. Acoustic Soc. Amer.* 25: 975–979, 1953.
5. Cooke, M.P, Brown, G.J. Crawford, M. and Green, P. "Computational Auditory Scene Analysis: listening to several things at once", *Endeavour*, 17(4): 186–190, 1993.
6. Hansen, J.H.L. Mammone, R.J. and Young, S. Editorial for the special issue of the IEEE Trans. SAP on robust speech processing". *IEEE Trans. SAP* 2(4): 549-550, 1994.
7. Green, P.D. Cooke, M.P. and M.D. Crawford, M.P. "Auditory Scene Analysis and Hidden Markov Model Recognition of Speech in Noise", *Proc. of ICASSP-95*, Vol.1: 401–404.
8. Kita, K., Kawabata, T. and Shikano, H. HMM continuous speech recognition using generalized LR parsing. *Trans. Info. Proc. Soc. Jap.* 31(3): 472–480, 1990.
9. Minami, Y. and Furui, S. "A Maximum Likelihood Procedure for A Universal Adaptation Method based on HMM Composition", *Proc. of ICASSP-95*, Vol.1: 129–132, 1995.
10. Nakatani, T., Okuno, H.G. and Kawabata, T. "Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System", *Proc. of AAAI-94*, 100–107, 1994.
11. Nakatani, T., Okuno, H.G. and Kawabata, T. "Unified Architecture for Auditory Scene Analysis and Spoken Language Processing", *Proc. of ICSLP-94*, 1403–1406, 1994.
12. Nakatani, T., Kawabata, T. and Okuno, H.G. "A computational model of sound stream segregation with the multi-agent paradigm", *Proc. of ICASSP-95*, Vol.4: 2671–2674.
13. Nakatani, T., Okuno, H.G. and Kawabata, T. "Residue-driven architecture for Computational Auditory Scene Analysis", *Proc. of IJCAI-95*, Vol.1:165–172, 1995.
14. Nakatani, T., Goto, M. and Okuno, H.G. "Localization by harmonic structure and its application to harmonic sound stream segregation", *Proc. of ICASSP-96*, Vol II:653–656, IEEE, 1996.
15. Okuno, H.G. ,Nakatani, T. and Kawabata, T. "Cocktail-Party Effect with Computational Auditory Scene Analysis — Preliminary Report —", *Symbiosis of Human and Artifact — Proc. of HCI Int'l '95*, Vol.2: 503–508, Elsevier Sci. B.V.
16. Okuno, H.G., Nakatani, T. and Kawabata, T. "Interfacing Sound Stream Segregation to Speech Recognition Systems — Preliminary Results of Listening to Several Things at the Same Time", *Proc. of AAAI-96*, *to appear*, 1996.
17. Ramalingam, C.S. and Kumaresan, R. "Voiced-speech analysis based on the residual interfering signal canceler (RISC) algorithm", *Proc. of ICASSP-94*, Vol.I: 473–476, 1994.
18. Rosenthal, D. and Okuno, H.G (Eds.). *Readings in Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates. Forthcoming.
19. Stadler, R.W. and Rabinowitz, W.M. "On the potential of fixed arrays for hearing aids", *J. of Acoustic Soc. Amer.* 94(3) Pt.1:1332–1342, 1993.
20. Warren, R.M. "Perceptual restoration of missing speech sounds", *Science*167: 392–393, 1970.