JAIST Repository
https://dspace.jaist.ac.jp/

| | |
|---|---|
| Title | Privacy protection for speech based on concepts of auditory scene analysis |
| Author(s) | AKAGI, Masato; IRIE, Yoshihiro |
| Citation | Proc. INTERNOISE 2012: 485 |
| Issue Date | 2012-08-22 |
| Type | Conference Paper |
| Text version | publisher |
| URL | http://hdl.handle.net/10119/10730 |
| Rights | Akagi, M. and Irie, Y. (2012/08/22). “Privacy protection for speech based on concepts of auditory scene analysis,” Proc. INTERNOISE2012, New York, 485. |
| Description | |

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

inter.noise 2012
august 19-22
new york city, usa
Quieting the World's Cities

# Privacy protection for speech based on concepts of auditory scene analysis

Masato AKAGI[a)]
School of Information Science, JAIST
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan

Yoshihiro IRIE[b)]
New Business Promotion Division, GLORY LTD.
1-3-1 Shimoteno, Himeji, Hyogo, 670-8567 Japan

**There are many sounds in the environment in which we live. In order to obtain the relevant information from a particular sound, we need to identify a target sound and extract the information from that. Being able to pinpoint a target sound from among two or more sounds is called the "cocktail party effect". The main component of this effect is auditory scene analysis (ASA), which, if it is working well, enables us to focus on something as soft as a whisper uttered several meters ahead. However, there is also the possibility that private conversations—e.g., those that take place at a bank, hospital outpatient department, pharmacy, etc.— may be overheard by another person. In this paper, we propose a method for protecting speech based on circumventing ASA. The proposed method renders the content of an utterance indistinct by generating an acoustical environment in which ASA does not work effectively. It does this by using an information masking technique to reproduce a sound that obscures the phonemes bearing the language semantic information of the conversations.**

## 1 INTRODUCTION

In public spaces such as banks, hospital outpatient departments, and pharmacies, it is easy for private conversations to be overheard, thus revealing secret or confidential information related to bank account balance, medical condition, etc. to strangers. It can also be annoying to accidentally overhear such information [1][2].

In general, the most effective method to prevent conversational speech from leaking is to keep speech sounds indoors. If people use a soundproof wall or room, they are isolated from outside parties and their conversation is secure against leaking. However, such a method cannot be used in open or public spaces. The next best method is to physically erase speech sounds:

[a)] email: akagi@jaist.ac.jp
[b)] email: y.irie@mail.glory.co.jp

active noise control (ANC), for example, can be used to erase speech sounds in all areas except where users are talking. However, this method is not always realistic considering the area within the range of effective erasing.

Another method currently being studied involves use of psychological techniques such as power masking (see Fig. 1) to prevent the perception of target speech [3][4]. Using such techniques has some drawbacks, however. The most significant is that in order to ensure privacy protection we have to raise the level of the mask sound, which we feel it is too noisy. The solution is to lower the level of the masking sound, but this makes it possible for outsiders to hear the conversation. For optimum speech privacy protection, it is necessary to keep the level of the power masking sound high to some extent, even if we feel it is too noisy. This, however, makes it difficult to use power masking in banks, pharmacies, etc.

Our proposed method for protecting speech privacy [5] is based on knowledge of the *perceptual fusion of sound*. In this method, a secondary sound (sound for hearing protection (SHP)) that obscures phonemes is reproduced simultaneously with the conversation. Outsiders hear the conversation as a different sound and feel that the speaker is talking about something they actually are not. Speech privacy protection is thus ensured. The central component of this method is really a type of information masking [6]. In the remainder of this paper, we give an overview of the basic concept and algorithm of the proposed method.

## 2 PERCEPTUAL FUSION OF SOUND: KNOWLEDGE FROM AUDITORY SCENE ANALYSIS

### 2.1 Auditory Scene Analysis

There are many sounds in the environment in which we live. In order to obtain the relevant information from a particular sound, we need to identify a target sound and extract the information from that. Being able to identify the target sound from among two or more sounds is called the “cocktail party effect”. The main component of this effect is auditory scene analysis (ASA) [7].

Bregman, considered one of the pioneers of ASA, was the first to discuss how we identify sounds when there are two or more sounds in the same environment. We briefly summarize some of his main points below.

When individuals hear music or speech, the first thing they do is disjointedly divide acoustic features into their primitives on the basis of the difference between the sound’s beginning, the difference of the sound’s arrival direction, pitch difference, timbre difference, etc. (separation). Next, they gather specific primitives that have the same properties using verbal knowledge, experience, etc. (grouping). Finally, they arrange the primitives to connect smoothly on the time axis (formation of the sound stream). In other words, they hear music or speech in three stages. We call these stages a series of work segregation. Segregation is the formation of meaningful information by not only separating sounds with different attributes but also grouping them by coming back to the structure of the attribute.

The rules for grouping the primitives and forming the sound stream have been reported by Bregman [8]. They are four psychoacoustically heuristic regularities (arranged as “rules” below) based on the properties of the sound coming from one sound source.

Rule (1): Common onset/offset
Rule (2): Harmonicity
Rule (3): Gradualness of change (smoothness)
Rule (4): Common changes occurring in the acoustic events

Even if two sounds exist at the same time, they do not start simultaneously, they do not share common fundamental frequencies and harmonics, the sound pressure rapidly changes when one sound starts, and they do not share common fluctuations of amplitude envelopes. Thus, individuals can use Rules (1)–(4) to easily extract and perceive the target sound as one sound stream passing through the processes—that is, the separation, grouping, and formation of the sound stream.

### 2.2 Speech privacy protection based on perceptual fusion

Here, we consider two sounds that are not segregated and therefore discerned as one sound. We call this situation perceptual fusion, which is essentially when all sounds fuse into one stream and are perceived as one sound even though there may be two sounds physically present. In this case, it becomes impossible to segregate speech correctly and the perceived sound has different information than the original. This results in listeners being unable to acquire verbal information from the perceived sound and speech privacy protection is thus achieved. In this situation, knowledge of segregation is applied oppositely.

### 2.3 Acceleration of perceptual fusion

To accelerate perceptual fusion for the purpose of protecting speech privacy, sound for hearing protection (SHP) must be perceptually fused with conversational speech. To do this, the primitives analyzed from conversational speech and SHP have to be collected and grouped into one sound. To put it more simply, SHP needs to ensure that all sounds are perceived as one sound according to the four regularities laid out by Bregman.

In our approach to SHP, we focus mostly on Rules (1) (common onset/offset), (2) (harmonicity), and (3) (gradualness of change (smoothness)). We also consider Rule (4) (common changes occurring in the acoustic events) and maintain speaker individuality, so that listeners do not feel any incongruity within SHP.

## 3 PROPOSED SYSTEM FOR SPEECH PRIVACY PROTECTION

In this section, we present our system for speech privacy protection based on the auditory scene analysis (ASA) concept discussed in section 2. A dataflow graph and output of each process is shown in Fig. 2.

The algorithm for calculating SHP is as follows.

1. A conversational speech wave captured with a microphone is converted into an amplitude spectrum (Fig. 2(a)) and the phase spectrum by FFT in each frame.
2. After the logarithmic transformation of the amplitude spectrum, the amplitude cepstra are calculated by IFFT.
3. The amplitude cepstra separates low frequency (spectrum envelope: phoneme information shown in Fig. 2(b)) and high frequency (spectrum fine structure: fundamental frequency and harmonics shown in Fig. 2(c)) by liftering.
4. The spectrum envelope converted from the low-liftered amplitude cepstra by FFT is modified as shown in Fig. 2(d). This modification is done by reversing the up and down envelopes on the boundary of the axis that preserves the average of the log spectrum. This guarantees that Rules (1) and (3) are followed.
5. The logarithmic amplitude spectrum (Fig. 2(e)) is obtained by combining the spectrum fine structure and the modified spectrum envelope. This guarantees that Rule (2) is followed.

6. The high frequency region of the amplitude spectrum is replaced with the original amplitude spectrum of the conversational speech (Fig. 2(f)). This guarantees that Rule (4)is followed and that speaker individuality is maintained.
7. The SHP wave is obtained by multiplying the modified amplitude spectrum by the preserved phase spectrum and then performing IFFT.

As shown in Fig. 3, the SHP wave is produced from the conversation speech acquired at the position of (1) by the above-mentioned algorithm and is presented by the speaker set at (2) between the talker and the listener. The conversation speech and SHP then come together to the listener at the position of (3). This causes the listener to perceptually fuse two sounds.

## 4 PERFORMANCE EVALUATION

We performed experiments to compare the protection performance of SHP with that of a regular pink noise, which has conventionally been the most effective sound to mask conversational speech. We determined articulation scores for words in cases in which one speaker presenting conversational speech was set up 3 m away from the listener and the other speaker presenting SHP was set up 1.5 m from the listener (Fig. 4). We created a word dataset to form a database for speech intelligibility testing using Japanese word lists produced by NTT-AT [9]. The sound pressure level for the conversational speech and the pink noise at listener’s ear was calibrated to 50 dB and that for SHP was varied between 44 and 56 dB. Listeners were asked to look at the speakers in front of them and to write what they heard on an answering sheet. All the listeners were JAIST graduate school students who did not have any hearing problems.

The experimental results are shown in Fig. 5. The horizontal axis indicates the three sound pressure levels of SHP and of the pink noise, and the vertical axis indicates the correct answer rate of the words. If we compare SHP at 50 dB with the pink noise, the correct answer rate is almost exactly the same. This demonstrates that the proposed method can offer the same degree of protection as the pink noise—which, as mentioned earlier, is said to be the best. A difference between the two is that, according to introspection reports of the listeners, SHP was not jarring whereas the pink noise felt annoying and largely incongruous. Listeners also felt that the conversational speeches were completely separate. When we slightly increased the sound pressure of SHP (to 56 dB), the correct answer rate decreased. Moreover, according to introspection reports of the listeners, it is clear that there is no sense of incongruity and annoyingness in this case.

Kishi et al. discussed unpleasantness when SHP is presented [10]. In this paper, unpleasantness of SHP is compared in listening tests with those of time-randomized and time-reversed reproduction speech and the pink noise as competitive masking noises. Figure 6 shows the result appeared in [10]. As shown in this figure, Type F (SHP) is the highest and type N (pink noise) is the lowest for the acceptable level, suggesting that with SHP, the unpleasantness is the smallest under high sound pressure.

## 5 SUMMARY

We presented a new method for speech privacy protection based on knowledge of ASA. This method provides SHP at the same time conversational speech occurs to obscure phoneme information and to generate sound in environments in which ASA cannot work very well (e.g., in open spaces). Experimental results showed that sounds perceived by the listener were different from the original sounds. A product using this method has already been released and is currently being introduced in Japanese pharmacies [11].

## 6 REFERENCES

1. Fujiwara, M., Hata, M., Yamakawa, T., and Shimizu, Y. (2011). “Experimental Study of Speech Privacy with a Sound-masking System in Pharmacies,” Proc. of Inter-noise 2011.
2. Yamakawa, T., Hata. M., Fujiwara, M., and Shimizu, Y. (2011). “The solution of speech privacy secured to a waiting space for a confidential conversation at a pharmacy counter,” Proc. of Inter-noise 2011.
3. Komiyama, T. and Kondo, K. (2011). “An efficient speech privacy system using speaker-dependent babble noise as maskers,” Proc. of Inter-noise 2011.
4. Ito, A. et al. (2007). “Oral information masking considering room environmental condition, Part 1: Synthesis of Maskers and examination on their masking efficiency,” Proc. of Inter-noise 2007
5. Akagi, M. and Irie, Y. (2011). “Privacy protection for speech based on concepts of auditory scene analysis,” IEICE Technical Report, EMM2011-59 (In Japanese).
6. Durlach, N. I. et al. (2003). “Note on information masking”, J. Acoust, Soc. Am, 113, 6, 2984-2987.
7. Bregman, A. S. (1990). “Auditory scene analysis: the perceptual organization of sound,” MIT Press, Cambridge, MA.
8. Bregman, A. S. (1993). “Auditory scene analysis: hearing in complex environments,” In Thinking in sound: the cognitive psychology of human audition (McAdams, S. and Bigand, E., Eds), Oxford Science Pub., Chapter 2.
9. Database for speech intelligibility testing using Japanese word lists, NTT-AT, March 2003.
10. Kishi, M., Morimoto, M., Sato, H., Kuroda, N., and Irie, Y. (2011). “The unpleasantness of masking noise of sound masking system,” Proc. 2011 ASJ fall meeting, 1-5-18 (In Japanese).
11. Glory speech privacy protection system:
 <http://www.glory-global.com/ir/pdf/k_110823e.pdf>

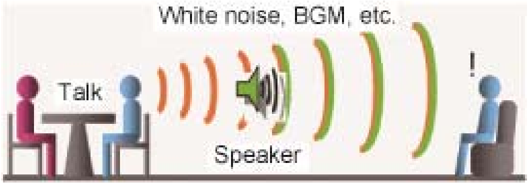

*Fig. 1 – Speech privacy protection by power masking with white noise, pink noise, BGM, etc.*

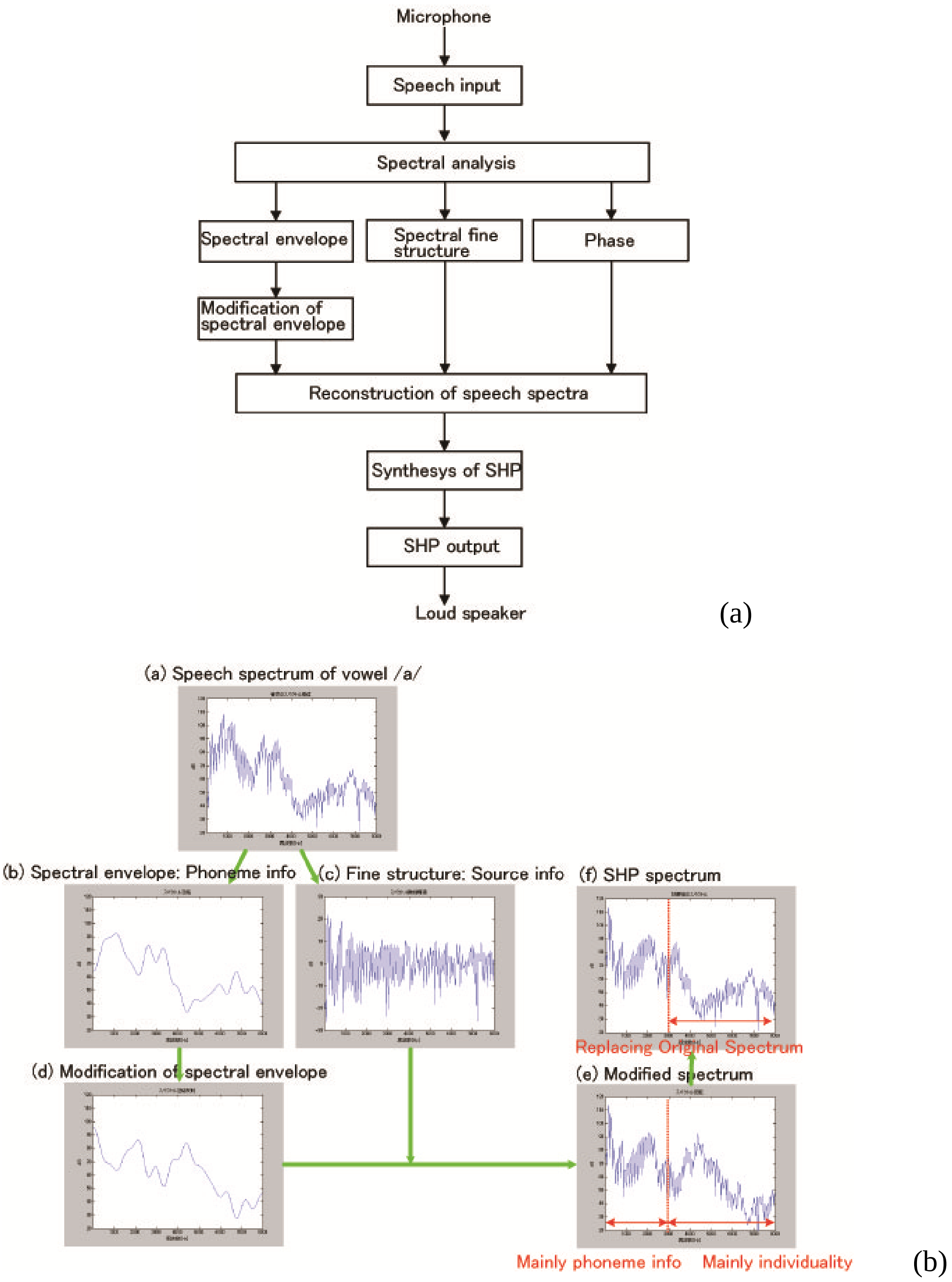
Microphone
Speech input
Spectral analysis
Spectral envelope
Spectral fine structure
Phase
Modification of spectral envelope
Reconstruction of speech spectra
Synthesys of SHP
SHP output
Loud speaker
(a)
(a) Speech spectrum of vowel /a/
(b) Spectral envelope: Phoneme info
(c) Fine structure: Source info
(f) SHP spectrum
Replacing Original Spectrum
(d) Modification of spectral envelope
(e) Modified spectrum
Mainly phoneme info
Mainly individuality
(b)

*Fig. 2 – (a) Dataflow and (b) resultant spectra of processes.*

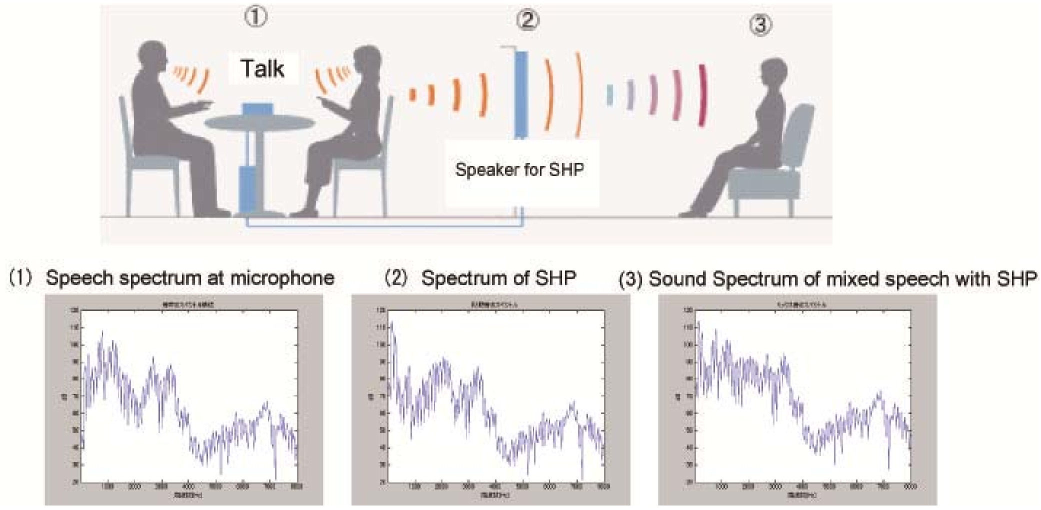
①
②
③
Talk
Speaker for SHP
(1) Speech spectrum at microphone
(2) Spectrum of SHP
(3) Sound Spectrum of mixed speech with SHP

*Fig. 3 – Proposed system for speech privacy protection.*

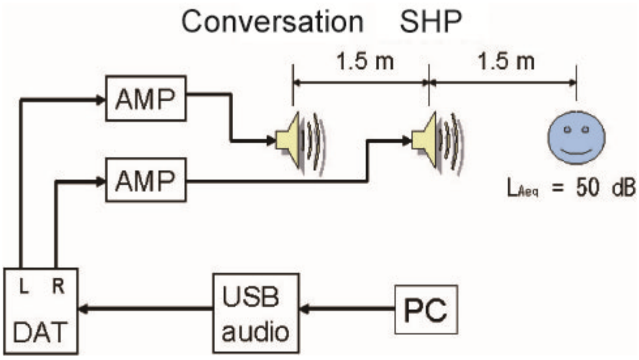

*Fig. 4 – Experimental setup.*

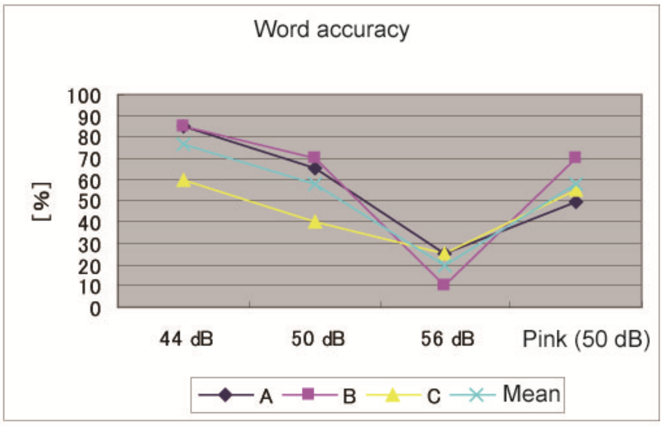

*Fig. 5 – Correct answer rate of the words.*

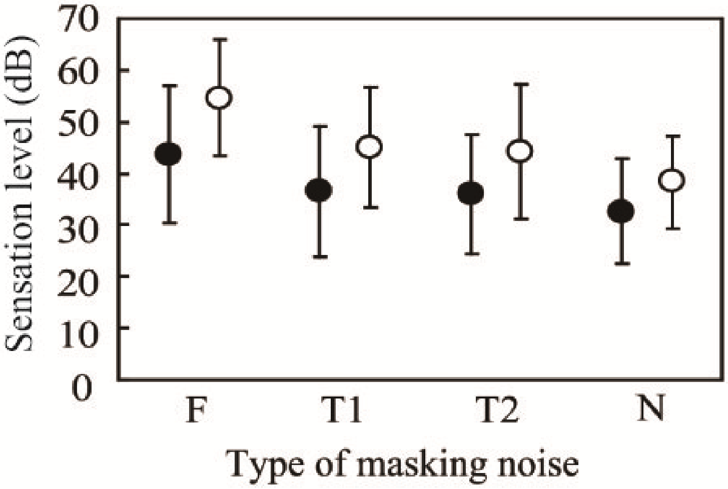

*Fig. 6 – Averaged acceptable levels of unpleasantness for each type of masking noise (Fig. 4 in Ref. [10]). F: SHP, N: pink noise, and T1/T2: time-randomized/time-reversed reproduction speech. The black and white circles indicate long-time and short-time situations, respectively. For the acceptable level, type F is the highest and type N is the lowest.*