EXTRACTING SPEECH FEATURES FROM HUMAN SPEECH LIKE NOISE

Daisuke KOBAYASHI, Shoji KAJITA, Kazuya TAKEDA and Fumitada ITAKURA

Graduate School of Engineering, Nagoya University
Furo-cho 1, Chikusa-ku, Nagoya, 464-01 JAPAN
takeda@nuee.nagoya-u.ac.jp

1. ABSTRACT

Human speech-like noise (HSLN) is a kind of bubble noise generated by superimposing independent speech signals typically more than one thousand times. Since the basic feature of HSLN varies from that of overlapped speech to stationary noise with keeping long time spectra in the same shape, we investigate perceptual discrimination of speech from stationary noise and its acoustic correlates using HSLN of various numbers of superposition. First we confirm the perceptual score, i.e. how much the HSLN sounds like stationary noise, and that the number of superposition of HSLN is proportional by subjective tests. Then, we show that the amplitude distribution of difference signal of HSLN approaches the Gaussian distribution from the Gamma distribution as the number of superposition increases. The other subjective test to perceive three HSLN of different dynamic characteristics clarifys that the temporal change of spectral envelope plays an important roll in discriminating speech from noise.

2. INTRODUCTION

Although detecting a human speech signal is one of the center functions of the human auditory system, automatic speech detection is still a difficult problem in various speech applications [1]. The most important issue in automatic speech detection is the robustness to the background noise or, in other words, how to design *discriminative* measure of speech from other sounds. From that point of view, it is needed to investigate not only the acoustic features of speech but also that of speech like noise.

We proposed Human Speech Like Noise (HSLN) as a bubble noise of more than one thousand times of superimposing [2] for the purpose of evaluation of noise robust spectral measures. HSLN has the important characteristic that by increasing the number of superposition, its basic feature changes from that of overlapped speech to stationary noise, although they both have similar long term spectra. Hence it is expected that we can find discriminative speech features by investigating HSLN.

The purposes of this paper are, thus, summarized as three points; 1) clarify that relationship between perceptual impressions of HSLN and the number of superposition, 2) find static and dynamic measures that correspond to the change of perceptual impressions, and 3) confirm the effectiveness of using above measures for speech detection.

In section 2, the results of a subjective test to find the critical number of superposition to discriminate overlapped speech from stationary noise will be described. In section 3 and 4, we will search for the physical measurement which highly correlates with perceptual impression of HSLN from the viewpoints of higher order statistics of amplitude distributions (section 3) and short time spectral change (section 4). Speech detection experiments combining static and dynamic measures are discussed in Section 5.

3. PERCEPTION OF HUMAN SPEECH LIKE NOISE

3.1. Generating HSLN

The HSLN used for tests are generated as follows.

1) Segment fifty sentences of ASJ phonetically balanced speech database of thirty male and thirty four female speakers into phrase utterances. Before the segmentation, each utterance is normalized so that the maximum amplitude is one. The normalization is executed in floating (`32 bit`) operation instead of the original ASJ speech database [3], which is recored in `16 bit` at `16 kHz` sampling frequency.

2): Randomly select a phrase speech chunk and concatenate it to form a long speech patch file.

3): Add N speech segments of L second length by folding the long speech patch file in every L second length. Superposition is done by adding N signals in float precision followed by a power normalization. Finally, HSLN are rounded to short (`16 bit`) data for D/A conversion.

3.2. Subjective Test

In order to investigate the critical number of superposition to discriminate overlapped speech from stationary noise, subjective tests were performed. In the tests, subjects selected the perceptual impression of the HSLN of different numbers of superposition from the four choices below.

Table 1: Conditions for the listening test.

# of subjects	10
# of superposition	2-4096
length of HSLN	1, 3, 10 sec
# of repetitions	3 / subject
DA condition	16 bit 16 kHz

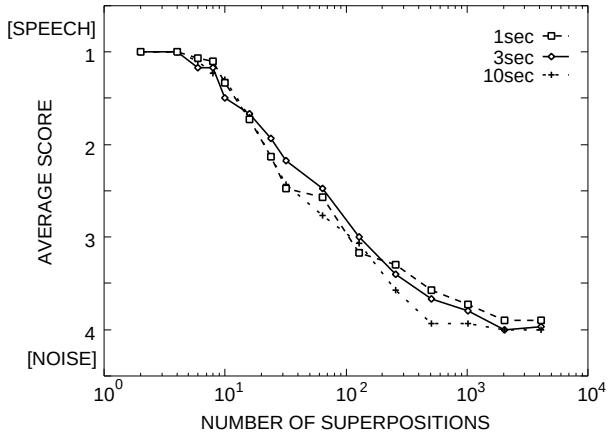

Figure 1: Averaged perceptual score across number of superposition of HSLN.

(1) sounds like overlapped speech.
(2) sounds like noise contaminated speech and speech is dominant.
(3) sounds like noise contaminated speech and noise is dominant.
(4) sounds like stationary noise.

The conditions of the tests are summarized in Table 1.

The obtained results illustrated in Figure 2 shows that the averaged perceptual score is proportional to the number of superposition. Although there is no discontinuity in changing the perceptual score, it can be observed that the HSLN of more than 256 times superposition is perceived as stationary noise and that of less than 64 times is perceived as overlapped speech.

4. HIGHER ORDER STATISTICS OF HSLN

It is well known that the amplitude distribution of speech follows Gamma distribution. In the case of HSLN of higher number superposition, however, amplitude distribution is expected to be a Gaussian because of the central limit thorem. It, therefore, makes sense that the higher order statistics of amplitude distribution of HSLN is a good cue of discriminating stationary noise from overlapped speech. As for the higher order statistics we calculated

$$\mathrm{Skew}(\boldsymbol{x}) = \left| \frac{1}{N} \sum_{j=1}^{N} \left[\frac{x_j - \overline{x}}{\sigma(\boldsymbol{x})} \right]^3 \right| \qquad (1)$$

and

$$\mathrm{Kurt}(\boldsymbol{x}) = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{x_j - \overline{x}}{\sigma(\boldsymbol{x})} \right]^4 - 3 \qquad (2)$$

for the HSLN of various numbers of superposition.

In Figure 2, Skew and Kurt of the amplitude distributions of HSLN and their difference signal of various superposition numbers are illustrated. The shaded area in the figures are indicating the 99 % confidence range of the hypothesis that the distribution follows a Gaussian distribution of N(0, 0.06). From the figures, it can be seen that kurtosis decrease to zero as the number of superposition increases which means that the amplitude distributions of HSLN and its difference signal approach the Gaussian distribution at higher superposition. The unevenness of statistics in lower numbers of superposition is greater in Skew than in Kurt in both signals. In the case of HSLN, Kurt is included in the confidence range where the number of superposition is greater than 64. On the other hand, in the case of its difference signal, Kurt is included in the confidence range where the number of superposition is greater than 256. The fact that the statics of the difference signal better corresponds to the subjective test suggests that the temporal structure plays an important role in perceiving speech.

5. TEMPORAL STRUCTURE OF HSLN

In this section, we will discuss the speech feature in HSLN from the standpoint of temporal structure. Since dynamic features of speech can be separated into dynamics of spectral envelope, fine structure of spectra and power, we remove one of the three dynamics from HSLN to test how its perceptual impression changes. Therefore the three test signals below are generated.

(1) The signal which obtained by removing the long term averaged spectral envelope from a HSLN. The residual signal of a 32 order LPC analysis of 3 seconds of HSLN is used as the test signal.

(2) The signal which is obtained by removing the short term spectral envelope from a HSLN. The test signal is a series of residual signals of 32 order LPC analyses on the windowed speech. The length of the window is 30 milliseconds, the period of the window is 10 milliseconds and the type of the window is Hamming.

(3) The signal which is obtained by removing the temporal change of power from a HSLN. This signal is calculated by normalizing the power of signal (1) every 10 milliseconds.

Using the above three test signals of HSLN of 32 to 256 superposition, the same subjective tests are performed as of Section 2. The obtained results are illustrated in Figure 3. In signal (1) and (3), the perceptual score is still proportional to the number of superposition. On the other hand, signal (2) is not perceived as human speech even in the lower number of superposition. From these results, it can be concluded that the speech features in HSLN

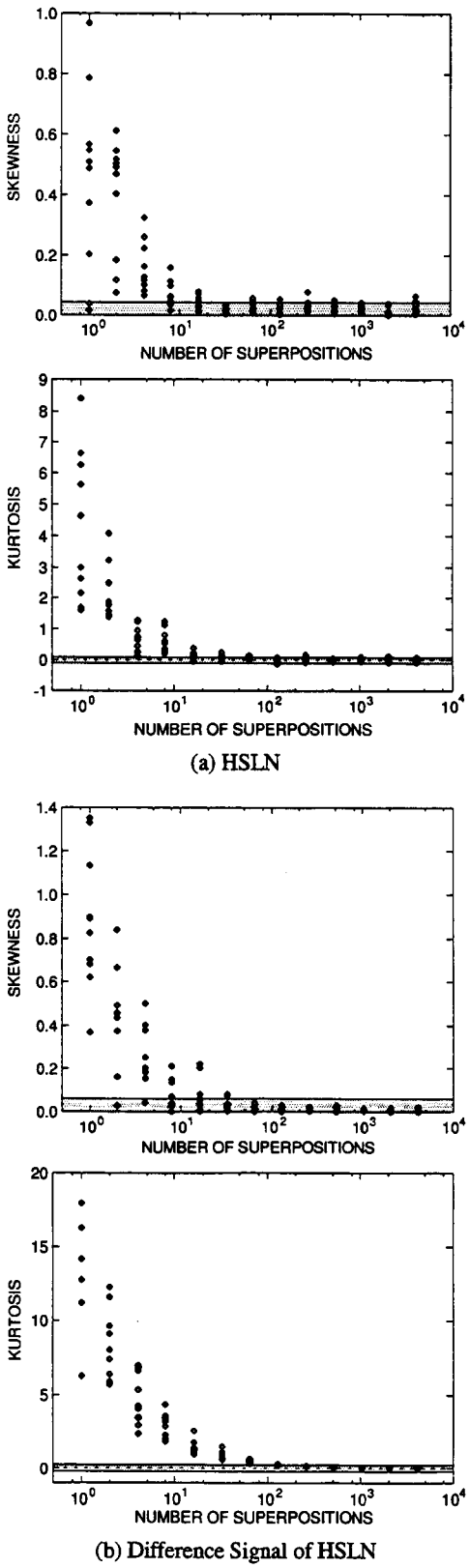

Figure 2: Skew and Kurt of HSLN and its difference signal.

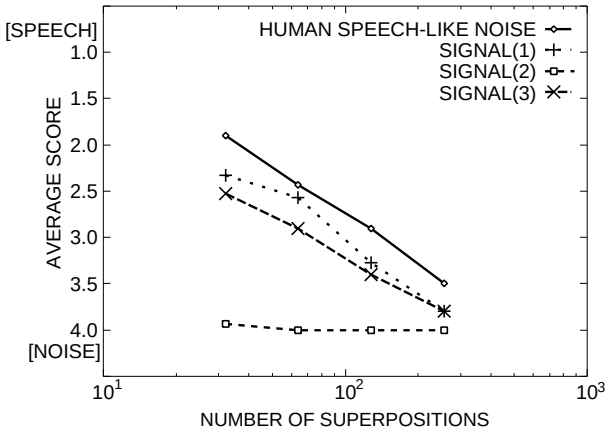

Figure 3: Comparison among HSLNs of different dynamic characteristics in correlates of perceptual score and number of superposition. (1) removing long term averaged spectral envelope; (2) removing short term spectral envelope; and (3) removing temporal change of power.

come from neither long term average spectral envelope nor temporal change of power and that the temporal change of the spectral envelope carries important speech features. The results obtained here support the literature that the dynamics of spectral structure is a good cue of detecting speech from other sounds, from the viewpoint of human perception.

6. SPEECH DETECTION EXPERIMENT

In the above two sections, we have shown that the two physical measurement Kurt of the difference signal and temporal change of the spectral envelope of HSLN describe the subjective listening test results. In this section, experiments of detecting speech segments from noisy speech are performed to confirm the effectiveness of combining those two measures for the speech detection problem. The experiments are designed as frame-by-frame decision if a speech signal exists or not.

We used norms of dynamic cepstrum parameters for an index of temporal change of spectral envelope;

$$D(t) = \sum_{l=-L}^{L} \sum_{n=0}^{q-1} \Delta c_n^2(t).$$

Thus combined measure of dynamic and static parameters are calculated by using weighting factor α

$$H(t) = \alpha \frac{D(t)}{W_D} + (1-\alpha) \frac{K(t)}{W_K}$$

where K(t) is Kurt of the difference signal, W_D and W_K are R.M.S. value of each measure for normalization.

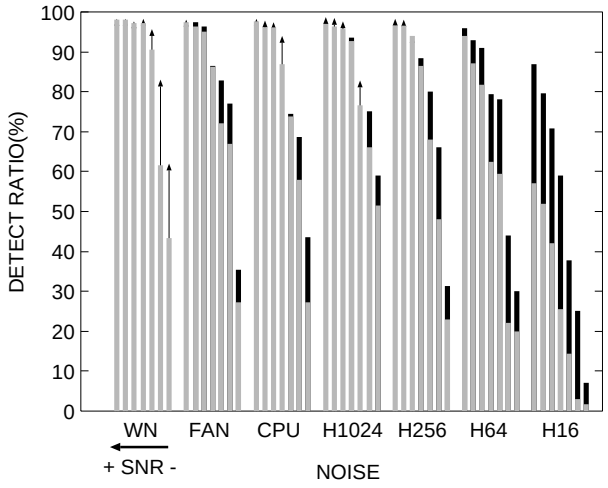

Figure 4: Results of speech detection experiment using combined measure of static and dynamic characteristics.

The original speech is a 53 second long utterance generated by concatenating randomly selected sentences in ASJ continuous speech database. Seven different noise signals (white Gaussian noise (`WN`), computer fan noise (`FAN`), computer room noise (`COM`), and HSLN of 1024 (`HSN1024`), 256 (`HSN256`), 64 (`HSN64`) and 16 (`HSN16`) superposition) are added to the original speech so that the global SNR of test signals are -10, -5, 0, 5, 10, 15, 20 dB for each noise signal.

Two different frames are used for calculating static and dynamic index. A three second long rectangular window is used for Kurt whereas a thirty millisecond long Hamming window is used for calculating 18 cepstrum coefficients through LPC analysis of order 16. The frame rate of both frames is ten milliseconds. The delta-cepstrum vectors are calculated using the adjacent five frames and the norms of the vectors are averaged over 61 frames to obtain the dynamic parameter $D(t)$.

Since the number of detected frames changes up to the detection threshold, performance in detecting speech is evaluated by the detection rate (the ratio of the number of correctly detected speech frames and the number of the total frames detected as speech) when the identification rate (the ratio of the number of correctly detected speech frames and the number of speech frames included in the test signal) equals to 95 %.

The obtained results are illustrated in Figure 4. Seven bars for seven different noise conditions indicate the detection rate at the different SNR conditions (25 dB to -10 dB, left to right) of the noise sound. Shadowed bars indicate the detection rate using dynamic measure only and black bars indicate the improvement of combining static measure. Upper arrows indicate the degradation due to the combination.

The improvement of detection accuracy by combining static and dynamic parameters is remarkable where the detection accuracy of dynamic parameter is low, especially the background noise is HSLN of fewer numbers of superposition. As shown in Figure 2, Kurt of the difference signal of 16 superposition HSLN is significantly lower than that of a single speech, i.e. no superposition. Therefore combining it with the dynamic parameter helped to discriminate speech from background bubble noise which can not deal with the dynamic parameter only.

7. SUMMARY

HSLN is investigated in order to find a physical measurement inherent feature of a speech signal. As a static parameter, Kurt of the difference signal of HSLN is found to be related with perceptual score. The importance of temporal change of the spectral envelope is also found by the other subjective test. The performance of speech detection is improved by combing static parameters (Kurt of the difference signal) with dynamic parameters (averaged norm of delta-cepstrum). The importance of both feature is also confirmed from the results of the speech detection experiment.

8. HSLN SAMPLES

In the ICSLP 96 CD-ROM, we put the human speech-like noise used in this study. The file names of the HSLNs from 1 to 4096 superposition are shown in Table 2.

Table 2: Samples of the HSN.

N	file name
1	SOUND A832S01.WAV
2	SOUND A832S02.WAV
4	SOUND A832S03.WAV
8	SOUND A832S04.WAV
16	SOUND A832S05.WAV
32	SOUND A832S06.WAV
128	SOUND A832S07.WAV
256	SOUND A832S08.WAV
512	SOUND A832S09.WAV
1024	SOUND A832S10.WAV
4096	SOUND A832S11.WAV

9. REFERENCES

1. Rabiner, L. R. and Sambur, M. R.: “An algorithm for determining the endpoints of isolated utterances”, Bell Syst. Tech. J., **54**, pp. 297–315 (1975).
2. Kajita, S. and Itakura, F.: ”Robust Speech Feature Extraction Using SBCOR Analysis”, Proc. ICASSP’95, Vol. I, pp.421-424 (1995)
3. Kobayashi, T., Itahashi, S., Hayamizu, S. and Takezawa, T.: ”ASJ continuous speech corpus for research”, J. Acoust. Soc. Japan, Vol. 48, pp.888-893 (1992). in Japanese.