Regular Paper

Predicting the Degradation of Speech Recognition Performance from Sub-band Dynamic Ranges

Masato Kondo,† Kazuya Takeda†,†† and Fumitada Itakura†,††

An acoustic measure for predicting the degradation of speech recognition performance due to noise contamination is developed. The merits of the proposed measure over using conventional SNR are that 1) the measure does not require original clean signal as a reference signal, 2) the measure takes the spectral shape of noise into account and, 3) the measure can be used to predict recognition performance directly. The basic idea of the measure is to utilize the dynamic range of the sub-band signals as an estimate of the SNR and to predict the degradation of recognition performance by taking the product of the recognition accuracy of each sub-band. The proposed measure is tested through experimental evaluation using white Gaussian noise and human-speech-like noise (HSN). In the experiment, the correlation between the predicted and the actual recognition accuracies are 0.96 and 0.99 for white noise and HSN, respectively. The results confirm the effectiveness of the proposed measure.

1. Introduction

In order to extend the speech recognition technology to a wider range of applications, the better robustness to changes of the environment is indispensable. In particular, technologies for dealing with noisy speech have been one of the most important issues in speech recognition research, and various approaches have been proposed[1]. One reason that so many different approaches have been attempted is that the 'noisy' condition is not a simple situation but varies greatly and therefore, evaluating the effectiveness of noisy speech recognition technologies is not easy.

The simplest way of comparing the noise reduction technologies under a given noise condition is to perform speech recognition experiments, because the noise condition of interest is not always the same as that under which the proposed method is developed. However, implementing and executing proper size of experiments of the reported noise reduction methods require much resources. One way to deal with this difficulty is to develop a measure which can be used to quantify the noise condition so that one can determine how much the noise condition under which the reported method functions well, is similar to the condition under consideration.

In many cases, the degradation of speech quality due to the existence of noise is characterized by the SNR (signal-to-noise ratio); however, utilizing SNR for that measure gives rise to two problems. The first problem is that calculating SNR essentially requires both an original clean speech and noise signals. In real situations, however, explicit detection of the noise segment is difficult[2),3)]. The second problem is that the SNR does not take into account the spectral shape of noise. In the case of in-car noise, for example, since the power of the noise signal is concentrated at lower frequencies (typically below 300 Hz), the SNR after eliminating the frequency region is of interest in most speech recognition applications. In general, the speech quality in higher-frequency region has less of an effect on the degradation of recognition performance. The SDR (signal deviation ratio) measure for various kinds of speech distortion has the same problem as the SNR measure.

In this paper, we propose an acoustic measure of speech degradation that can directly predict the speech recognition performance based on dynamic ranges of the signal in various frequency regions (sub-band dynamic ranges). The proposed method involves two ideas. The first idea is that the spectral shape of the noise can be measured without reference, by calculating the sub-band dynamic range (**Fig. 1**). Since the low-energy segment, e.g., silence, of the signal is masked by background noise, dynamic range is a reasonable quality index of noisy signals. Thus, by dividing the noisy signal into several frequency bands and by cal-

† Graduate School of Engineering, Nagoya University
†† Center for Integrated Acoustic Information Research, Nagoya University

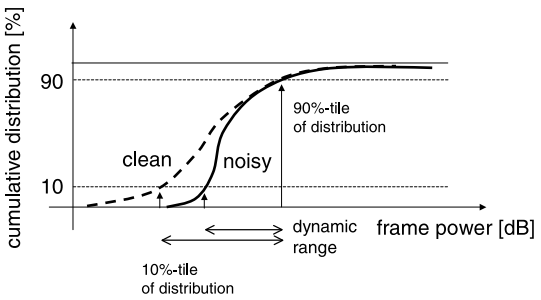

Fig. 1 Cumulative distribution of frame power and the dynamic range of a signal. When the noise signal masks the speech signal, the lower bound of the distribution of frame powers becomes higher.

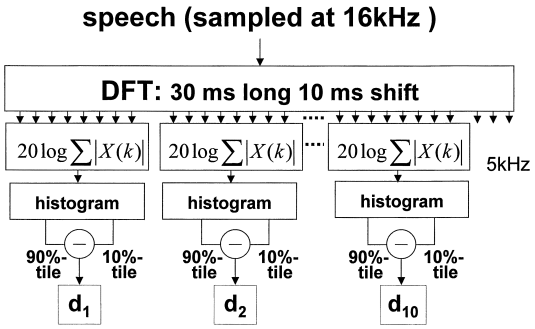

Fig. 2 Calculation method for sub-band dynamic range.

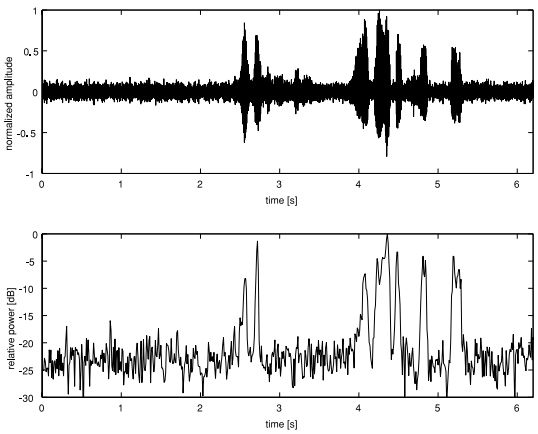

Fig. 3 An example of the log frame power sequence of a sub-band (3.0–3.5 kHz) when human speech like noise of 1,024 superpositions is added with the global SNR of 5 dB. Full-band waveform (upper) and sub-band log frame power sequence (lower).

culating the dynamic range of each sub-band signal, the spectral shape of the noise can be estimated. The second idea is that the effects of the different band-limited noises on the speech recognition performance are independent of each other[4)~7)]. Considering this idea, it is assumed that the recognition performance can be calculated as the product of the estimated accuracies, each of which is determined by the SNR of a particular frequency band.

The remainder of this paper consists of the following four sections. In Section 2, the method for calculating the sub-band dynamic range is described. In Section 3, recognition experiments for determining the degradation of recognition accuracy as a function of sub-band dynamic range are detailed. In Section 4, expected recognition errors due to degradation in all sub-bands are integrated to form a prediction measure of recognition performance. Experimental results of the evaluation tests are also described. A summary of this paper will be given in Section 5.

2. Sub-band Dynamic Range Calculation

In this paper, sub-band dynamic range values are calculated as shown in **Fig. 2**. All speech data are digitized into 16 bits at 16 kHz sampling rate. Short-time Fourier analysis is executed on a 30 ms frame at intervals of 10 ms. Because of the relative importance of the lower frequency region in speech recognition, the frequency range of 0 to 5 kHz is used for calculating sub-band dynamic ranges. The spectral component of the frequency range of 0 to 5 kHz is divided into 10 equal-width sub-bands, i.e., 0–500, 500–1000, ..., 4.5 k–5.0 kHz. Then the log-power of the sub-band in the time frame is calculated. Finally, after taking the histograms of sub-band frame powers over the given utterances, the sub-band dynamic ranges are calculated as the differences between 90%-tile and 10%-tile values of the histograms.

Under certain conditions, this sub-band dynamic range describes the ratio of the powers of background noise and the speech signal. Suppose that the average frame power of background noise is lower than that of the speech signal, i.e., the positive SNR condition, as shown in **Fig. 3**. In this case, the distribution of the log frame power is expected to exhibit two-peaks (one is of the noise and the other one is of the speech signal) and the distance between the two peaks is the average SNR.

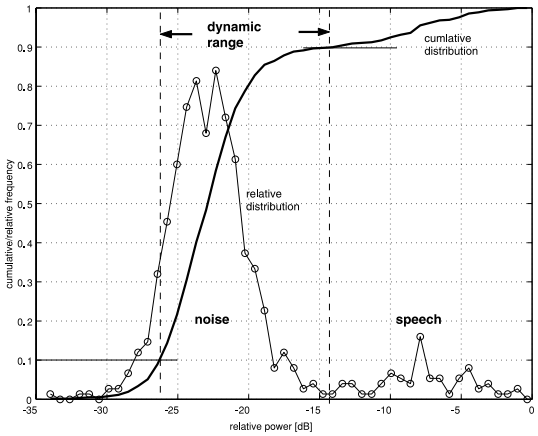

Fig. 4 Histogram and cumulative distribution of the sub-band log frame powers of the noisy speech given in Fig. 3.

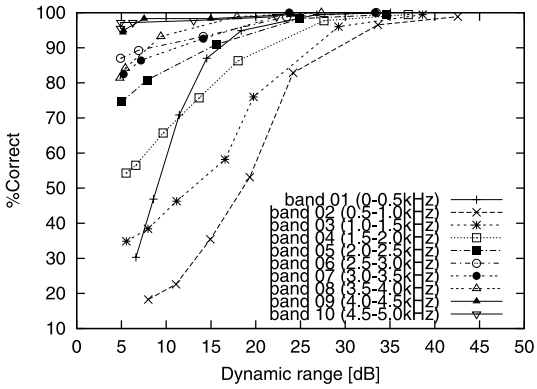

Fig. 5 Recognition performance as a function of the sub-band dynamic range.

In the above example, the density function of the log frame power has the shape given in **Fig. 4**. In that figure, it is clear that the lower tail of the distribution corresponds to the noise signal, whereas the higher tail corresponds to the speech signal. Thus, although it is difficult to estimate the mean from distributions, the characteristics of the two distributions can be estimated from the lower and higher percentile values of the frame power distribution. In the case of Figure 4, the 10%-tile value of the overall distribution is about 5 dB below the mean of the noise frame power, whereas the 90%-tile value is about 7 dB below the mean of speech frame power. The difference between them is a reasonable estimate of the SNR. The reader should note that the lower and higher endpoints used to calculate dynamic range used in this paper, i.e., 10% to 90%, are reasonable only when the number of samples (frames) of noise and speech signals are comparable.

3. Sub-band Dynamic Range and Recognition Performance

It is well known that the phonetic feature of speech signals is mainly contained in the frequency range lower than 3 kHz, and that the energy of the speech signal is concentrated at lower frequencies with about −6 dB/Oct of spectral tilt. Thus, the degradation of recognition performance due to noise is a function of not only the total power of the noise but also its spectral shape. In this section, we will determine the relationship between recognition accuracy and the SNR in particular frequency bands, through recognition experiments.

3.1 Recognition Experiments

The experimental conditions are as follows. As speech material, thirty sentences spoken by a male speaker are extracted from the JNAS newspaper corpus of Acoustic Society Japan[8]. The 1997 IPA standard monophone models (16 mixture, three-state, 43 phone-set, gender dependent) are used for acoustic modeling[9]. The feature vector for the experiment is 25 MFCC's (12 static + 12 delta + delta-log-power). The length and shift period of the analysis window are 25 ms and 10 ms, respectively. The recognition task material is dictation of 311 phrases without grammar. The index of the recognition performance is %Correct given by

$$\frac{N-S-D}{N},$$

where N, D and S are the total numbers of words, deletion errors and substitution errors, respectively.

3.2 Noise in Single Sub-band

In the first experiment, the effect of noise of a particular frequency band on recognition accuracy is examined. In this experiment, band-limited white noise is added to the original utterances. The band of the noise is changed from 0–0.5 kHz to 4.5–5.0 kHz, as are the overall SNR conditions (5–40 dB). The results of the recognition experiments are plotted in **Fig. 5**. In the figure, the recognition score is plotted as a function of the SNR measured in the sub-band dynamic range.

It can be seen in all sub-bands that the recognition performance decreases as the dynamic range of the band decreases. The performance

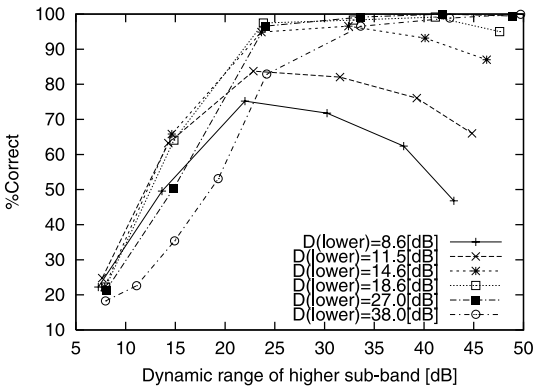

Fig. 6 Recognition performance as a function of the sub-band dynamic range of 0.5–1.0 kHz, i.e., the higher sub-band, when noise also exists in 0–0.5 kHz, i.e., the lower sub-band, with SNR of D (lower).

more strongly depends on the sub-band dynamic range in the lower frequency region. For the recognition system, a dynamic range value of more than 25 dB is needed in the sub-band of 0.5–1 kHz, to obtain a recognition accuracy of more than 80% in this particular task. Dynamic ranges of more than 20 dB and 15 dB are also needed for 1–1.5 kHz and 1.5–2.0 kHz bands, respectively, for an 80% score. On the other hand, the change of the dynamic range in sub-bands higher than 3.5 kHz does not affect the recognition performance, unlike in the lower bands.

3.3 Adding Noise to Multiple Bands

In **Fig. 6**, recognition performance is plotted as a function of the dynamic range of 0.5–1.0 kHz band. Unlike in the previous section, in this case, a different band-limited noise is added simultaneously to 0–0.5 kHz. Thus, in this experiment, the dependency of neighboring frequency bands in affecting the recognition accuracy is examined.

As shown in the figure, unlike the previous result, the relationship between the sub-band dynamic range and recognition performance is not monotonic. When the dynamic range of the lower (0–0.5 kHz) band is less than 10 dB, for example, the recognition performance decreases as the dynamic range of the higher (0.5–1.0 kHz) band increase to above 25 dB. This result suggests that the great difference in SNR conditions between neighboring frequency regions causes severe degradation in recognition performance, particularly in the lower frequency regions. Thus, as discussed later, in integrating sub-band information to predict recognition accuracy, assuming independence between neighboring frequency bands in the lower frequency region is inappropriate. In experiments, such a non-monotonic relationship, however, is not found in any other combination of sub-bands.

4. Predicting Recognition Performance

4.1 Prediction Formula

From the previous results, the recognition performance is determined to be a function of the sub-band dynamic range. In this section, by integrating the recognition performance, which is a function of the dynamic range of sub-bands, an acoustic measure is constructed to predict the recognition performance for noisy speech. As for the integrating principle of sub-band information, a simple product of the recognition accuracy in each band is adopted:

$$(1-e) = \prod_{i=1}^{10} (1 - e_i(d_i)), \qquad (1)$$

where, e_i stands for the error rate due to the noise in the ith sub-band. The error rate is given as a function of the dynamic range of the ith sub-band d_i.

It should be noted that this is the inverse form of Fletcher's definition of intelligibility[4),5)]. In our case, "the total degradation of the recognition performance is governed by the accumulation of recognition performance in each sub-band", whereas in Fletcher's case "the total intelligibility is controlled by the accumulation of the degradation of intelligibility in each sub-band", i.e.,

$$(1-s) = \prod (1 - s_i),$$

where s and s_i are intelligibilities of full band and ith sub-band, respectively.

It should also be noted that $e_i(d_i)$ is assumed to be a ratio in the period of $(0, 1)$. In other words, the 'word accuracy' score that takes word insertion error into account cannot be predicted by this idea, because the score can take a negative value, particularly under poor conditions.

Furthermore, from the results in the previous section, dependence between the lower two sub-bands should be taken into account. In this case, the prediction formula can be rewritten as

$$(1-e) = \{1 - e_{1,2}(d_1, d_2)\} \prod_{i=3}^{10} \{1 - e_i(d_i)\}. \quad (2)$$

Finally, the expected recognition performance is normalized by the baseline performance of the system:

$$P_{\text{pred}} = P_0 \frac{1 - e_{1,2}(d_1, d_2)}{1 - e_{1,2}(\infty, \infty)} \prod_{i=3}^{10} \frac{1 - e_i(d_i)}{1 - e_i(\infty)}, \quad (3)$$

where P_0 is the baseline performance of the recognizer at the given task, i.e., the recognition accuracy of the clean speech case. $e_i(\infty)$ is the word error rate when no noise signal is contaminated in the ith sub-band. Therefore, $1 - e_i(\infty) = P_0$ holds for each band.

4.2 Evaluation of the Prediction Form

In this section, the proposed prediction form is evaluated through recognition experiments. In the experiments, the experimental results in Sections 3.1 and 3.2 are used as the error rate functions, $e_i(d_i)$. As for the test utterances, 20 different sentences spoken by the same speaker as in the previous section are used. Noise signals are added to the test utterance, then recognition experiments are performed in the same manner as described in Section 3. Thus, the performed experiments are 'closed' with respect to the speaker, the recognition system and the recognition task conditions. The predicted recognition performance is calculated from sub-band dynamic ranges $\{d_i\}$ according to Eq. (3).

As test noise, a set of white Gaussian noise (WGN) and a set of human-speech-like noise (HSN)[10] are used. HSN is a kind of bubble noise generated by superimposing independent speech signals. When the number of superpositions is small, the signal simulates a multi-speaker situation, whereas when the number of superpositions increases to some hundreds, the HSN becomes stationary noise whose spectral shape represents long-term spectra of speech. In the experiments, three sets of HSN consisting of 32, 256 and 1024 superpositions are used. For WGN, 10 band-limited and one full band signal are prepared. These two sets of noise signals are added to the test utterance under 6 different SNR conditions, i.e., 0, 5, 10, 20, 30, 40 dB; therefore, 18 HSN and 66 WGN test conditions are prepared in total. Note that the SNR here is calculated from original speech sound and noise signals for each sentence. Sub-band dynamic ranges $\{d_i\}$, on the other hand, are calculated over given 20 sentences, by the method described in Section 2.

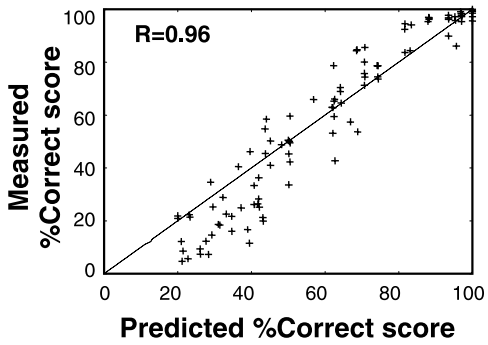

Fig. 7 Correspondence between predicted and measured performances (white Gaussian noise; SNR is estimated using 20 sentences).

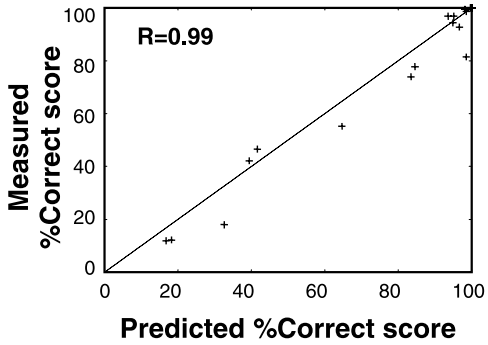

Fig. 8 Correspondence between predicted and measured performances (human speech like noise case).

4.3 Results of Experiments

The relationship between predicted and actual recognition performances is plotted in **Figs. 7** and **8** for WGN and HSN, respectively. In the figures, it can be seen that Eq. (3) can be used to predict the degradation of recognition performance due to a noise well. The correlation coefficients between actual and predicted recognition rates are 0.96 for WGN and 0.99 for HSN.

The same experiments as described in the previous section, but with changes of two conditions, i.e., 1) the length of the signal for estimating sub-band dynamic ranges and 2) disregard of the correlation between the lower two sub-bands in modeling, are performed. Under the first condition, the sub-band dynamic range was calculated from one sentence, whereas 20 sentences were used in the previous experiment. The average duration of the sentences

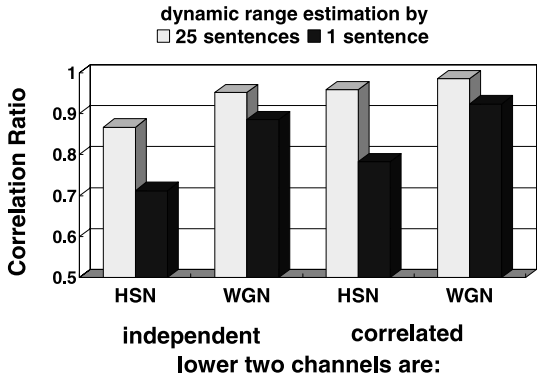

Fig. 9 Correlation coefficients between predicted and actual recognition accuracy calculated under various conditions (WGN: white Gaussian noise, HSN: Human-speech-like noise).

was about three seconds. For the second condition, the independence between the lower two frequency bands in error rate estimation, i.e.,

$$e_{12}(d_1, d_2) = e_1(d_1) \cdot e_2(d_2),$$

is assumed. The correlation coefficients of the predicted and actual recognition performance are given in **Fig. 9**, for all conditions.

The results clarified that accurate prediction is difficult when the sub-band dynamic range is calculated using only one sentence. It is also confirmed that taking the dependence between the lower two frequency bands into account is important.

5. Summary

A method of predicting the degradation in speech recognition performance due to noise is proposed. Since the prediction measure is defined based on the dynamic ranges of sub-bands, an original speech signal as reference is not needed, unlike in the case of conventional SNR measure. Furthermore, the measure can take the spectral shape of the noise into account when predicting recognition performance. Experimental evaluations were performed under the 'closed' condition where the same speaker, the same recognition system and the same recognition task under the trained condition, but different noise and speech signals were used. The experimental results, although obtained under limited conditions, confirmed the effectiveness of the measure. The correlation coefficients between predicted and actual recognition rates are 0.96 and 0.99 for white Gaussian noise and human-speech-like noise, respectively.

Since the proposed measure uses the error rate function, $e_i(d_i)$, for the specific recognition conditions, further research is needed to extends the method to a wider range of noise conditions. In particular, whether or not the same error rate function can be used for different speakers, i.e., speaker dependence, is the most important issue that was not clarified in this study.

Also, further research on the following points is expected to improve the performance of the measure: 1) a method of dividing the frequency band into sub-bands, 2) the effectiveness of weighting each sub-band output and 3) optimum amount of data for calculating dynamic range values.

Acknowledgments The authors are grateful to the reviewers who gave appropriate comments. This research has been supported by a Grant-in-Aid for COE Research (No.11CE2005).

References

1) Junqua, J-C. and Haton, J-P.: *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers (1996).
2) Hirsch, H.G. and Ehrlicher, C.: Noise Estimation Techniques for Robust Speech Recognition, *Proc. International Conference on Acoustic Speech and Signal Processing* (*ICASSP '95*), Detroit, pp.153–156 (1995).
3) Korhauer, A.: Robust Estimation of the SNR of Noisy Speech Signals for the Quality Evaluation of Speech Database, *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pp.123–126 (1999).
4) Fletcher, H.: *Speech and Hearing in Communication*, Allen, J.B. (Eds.), Acoustical Society of America, pp.278–302 (1995).
5) Allen, J.B.: How Do Human Process and Recognize Speech?, *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.4, pp.567–577 (1994).
6) Bourlard, H. and Dupont, S.: A new ASR approach based on independent processing and recombination of partial frequency bands, *Proc. International Conference on Spoken Language Processing,* (*ICSLP '96*), Philadelphia, pp.426–429 (1996).
7) Okawa, S., Eorico, B. and Potamianos, A.: Multi-band Speech Recognition in Noisy Environment, *Proc. International Conference on Acoustic Speech and Signal Processing* (*ICASSP '98*), Seattle, pp.641–644 (1988).
8) Ito, K. and Yamamoto, M., et al.: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition, *Journal*

of Acoustic Society Japan (*E*), Vol.20, No.3, pp.199–206 (1999).
9) Kawahara, T. and Lee, A., et al.: Japanese Dictation Toolkit —1997 version, *Journal of Acoustic Society Japan* (*E*), Vol.20, No.3, pp.233–239 (1999).
10) Kobayashi, D., Kajita, S., et al.: Extracting Speech Features from Human Speech-like Noise, *Proc. International Conference on Spoken Language Processing* (*ICSLP'96*), Philadelphia, pp.418–421 (Oct. 1996).

(Received November 16, 2001)
(Accepted April 16, 2002)

Masato Kondo received B.E.E. and M.E.E. from Nagoya University in 1999 and 2001 respectively. He had studied a measure for speech quality for his undergraduate research project at Nagoya University.

Kazuya Takeda received B.E.E. and M.E.E. and Doctor of Engineering degrees all from Nagoya University in 1983, 1985 and 1994, respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated a project for constructing voice-activated telephone extension system. Since 1995, he has been working for Nagoya University as an Associate Prof.

Fumitada Itakura was born in Toyokawa, Japan on August 6, 1940. He received the B.E.E., M.E.E., and Doctor of Engineering degrees all from Nagoya University in 1963, 1965 and 1972, respectively. In 1968, he joined the Electrical Communication Laboratory of NTT, Musashino, Tokyo, and participated in the speech processing, including the maximum likelihood spectrum estimation, the PARCOR method, the line spectrum pair method, the composite sinusoidal method, and the APC-AB speech coding. In 1981, he was appointed to Head of Speech and Acoustics Research Section of the ECL, NTT. In 1984, he left NTT to become a professor of Nagoya University, where he teaches courses of communication theory and signal processing. In 1975, he received IEEE ASSP Senior Award for his paper on speech recognition based on the minimum prediction residual principle. He is a co-recipient with B.S. Atal of 1986 Morris N. Liebmann Award for contributions to linear predictive coding for speech processing. In 1997, he received IEEE Signal Processing Society Award.