論文 / 著書情報
Article / Book Information

Title	Noise robust speech recognition using F0 contour information
Authors	Koji Iwano, Takahiro Seki, Sadaoki Furui
出典 / Citation	IEICE Transactions on Information and Systems, Vol. E87-D, No. 5, pp. 1102-1109
発行日 / Pub. date	2004, 5
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2004 Institute of Electronics, Information and Communication Engineers.

IEICE TRANS. INF. & SYST., VOL.E87-D, NO.5 MAY 2004

PAPER Special Section on Speech Dynamics by Ear, Eye, Mouth and Machine

Noise Robust Speech Recognition Using F_0 Contour Information

Koji IWANO[†a)], *Member*, Takahiro SEKI[†*], *Nonmember, and* Sadaoki FURUI[†], *Fellow*

SUMMARY This paper proposes a noise robust speech recognition method using prosodic information. In Japanese, the fundamental frequency (F_0) contour represents phrase intonation and word accent information. Consequently, it conveys information about prosodic phrases and word boundaries. This paper first describes a noise robust F_0 extraction method using the Hough transform, which achieves high extraction rates under various noise environments. Then it proposes a robust speech recognition method using multi-stream HMMs which model both segmental spectral and F_0 contour information. Speaker-independent experiments are conducted using connected digits uttered by 11 male speakers in various kinds of noise and SNR conditions. The recognition error rate is reduced in all noise conditions, and the best absolute improvement of digit accuracy is about 4.5%. This improvement is achieved by robust digit boundary detection using the prosodic information.

key words: *noise robust speech recognition, prosody, fundamental frequency (F_0) contour, multi-stream HMM, Hough transform*

1. Introduction

Recently, continuous speech recognition has made great progress and high recognition rates can be achieved for read speech uttered in a clean/quiet environment. However, current speech recognition technology has not matured to the point where it can provide high performance for spontaneous speech tasks or when used in a noisy environment.

Based on the importance of prosodic features in the human speech perception process, several experiments using prosodic features in automatic speech recognition process have been conducted [1]. For example, prosodic information has been used for improving the performance of phoneme recognition [2], [3] and for improving language models [4]. For spontaneous speech recognition, it has been reported that the recognition performance can be improved by predicting the appearance of spontaneous-speech specific events using prosodic features, such as pause length and syllable duration [5], since spontaneous speech includes many repairs and disfluencies, which decrease the recognition performance [6].

Since it has been found that human beings use prosodic information to increase the robustness in recognizing speech when acoustic information is unreliable [7], prosodic information should be useful and effective for noise robust automatic speech recognition. However, until the present, prosodic information has been hardly explored as a means to increase noise robustness in continuous speech recognition.

From this point of view, this paper proposes a noise robust speech recognition method using prosodic information. Our method uses fundamental frequency (F_0) contours. Since it represents phrase intonation and word accent in Japanese utterances, they are expected to be useful to detect phrases and word boundaries. For reliably detecting the F_0 values, the Hough transform, which is a robust image processing method, is used in the F_0 extraction step.

This paper is organized as follows: In Sect. 2, a robust F_0 extraction method using the Hough transform is described. Section 3 proposes our noise robust speech recognition method using multi-stream HMMs combining segmental and prosodic information. Experimental results are reported in Sect. 4, and Sect. 5 concludes this paper.

2. F_0 Extraction Using the Hough Transform

2.1 Hough Transform

The Hough transform is a technique to robustly extract parametric patterns, such as lines, circles, and ellipses, from a noisy image [8].

This paper uses the Hough transform method to extract linear transitional patterns of the F_0 values. The method for extracting a significant line from an image on the x-y plane can be formulated as follows. Suppose the image consists of n pixels at (x_i, y_i) $(i = 1, \cdots, n)$. Every pixel on the x-y plane is transformed to a line on the m-c plane as

$$c = -x_i m + y_i \quad (i = 1, \cdots, n) \tag{1}$$

A brightness value of the pixel on the x-y plane is accumulated at every point on the line. This process is called "voting" to the m-c plane. After voting of all pixels, the maximum accumulated voting value on the m-c plane is detected, and the peak point (m, c) is transformed to a line on the x-y plane by the following equation:

$$y = mx + c \tag{2}$$

Figure 1 shows an example of the Hough transform process. In this example, an image consists of five pixels on the x-y plane. Dotted arrows indicate the voting processes in the Hough transform, in which each pixel is transformed into a line on the m-c plane and the brightness values of all the pixels are accumulated at every point along these lines.

Manuscript received September 10, 2003.
Manuscript revised November 13, 2003.

[†]The authors are with the Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.
[*]Presently, with IBM Global Service-Japan.
a) E-mail: iwano@furui.cs.titech.ac.jp

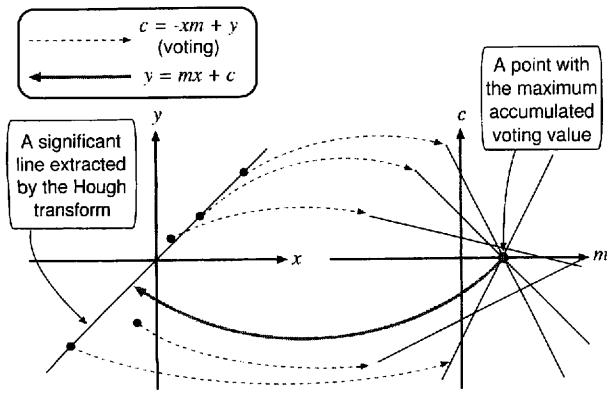


Fig. 1 An example of the Hough transform process.

A significant line on the x-y plane is obtained by transforming a point with the maximum voting value on the m-c plane as shown by the gray solid arrow in the figure.

2.2 F_0 Extraction Using the Hough Transform

Although F_0 contours have temporal continuity in the voiced period, the cepstral peaks which have been widely used to extract the F_0 values often cause errors, including half pitch, double pitch and drop outs, due to noise effects. To take advantage of the continuity, the Hough transform is applied to time-cepstrum images of noisy speech.

Speech waveforms are sampled at 16 kHz and transformed to 256 dimensional cepstra. A 32 ms-long Hamming window is used to extract frames every 10 ms. For reducing noise effects of a high frequency domain, we extract and use time-cepstrum images which are limited to 60–256 dimensions and liftered according to the following formula:

$$c'_d = \left\{0.6 + 0.4 \sin\left(\frac{d-60}{140-60} \times \frac{\pi}{2}\right)\right\} \cdot c_d \tag{3}$$

where c_d is the original dth cepstrum and c'_d is the liftered cepstrum. To the liftered time-cepstrum image, a nine-frame moving window is applied at every frame interval to extract an image for line information detection. The time-cepstrum image is used as the pixel brightness image for the Hough transform. An F_0 value is obtained from a cepstrum index of the center point for the detected line. Since the moving window has nine frames, the time continuity for 90 ms is taken into account in this method.

3. Integration of Segmental and Prosodic Information for Noise Robust Speech Recognition

3.1 Japanese Connected Digit Speech

The effectiveness of our method was evaluated in a Japanese connected digit speech recognition task. In Japanese connected digit speech, two or three consecutive digits usually make one prosodic phrase. Figure 2 shows an example of an F_0 contour of connected digit speech. The first two digits make the first prosodic phrase, and the latter three digits

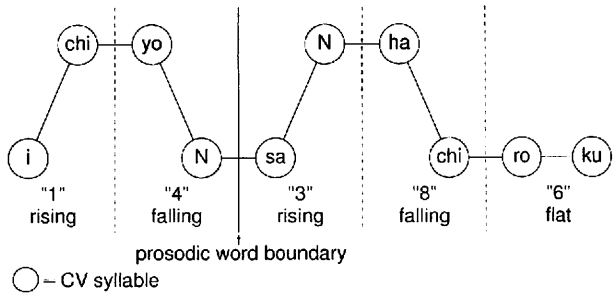


Fig. 2 An example of F_0 contour of Japanese connected digit speech.

make the second prosodic phrase. The transition of F_0 is represented by CV syllabic units, and each CV syllable can be prosodically labeled as a "rising", "falling", or "flat" F_0 part. Since this F_0 feature changes at digit boundaries, the accuracy of digit alignment in the recognition process is expected to be improved by this information.

3.2 Integration of Segmental and Prosodic Features

Each segmental feature vector has 25 elements consisting of 12 MFCC, their deltas, and the delta log energy. The window length is 25 ms and the frame interval is 10 ms. Cepstral mean subtraction (CMS) is applied to each utterance.

Two kinds of prosodic features are extracted; one is the $\Delta \log F_0$ value which represents the F_0 transition, and the other is the maximum accumulated voting value, denoted as "MAVV", obtained in the Hough transform which indicates the degree of temporal continuity in the F_0. Prosodic feature vectors consist of both or either of the two features.

$\Delta \log F_0$ value is calculated as follows:

$$\Delta \log F_0 = \frac{d \log F_0}{dt} \tag{4}$$

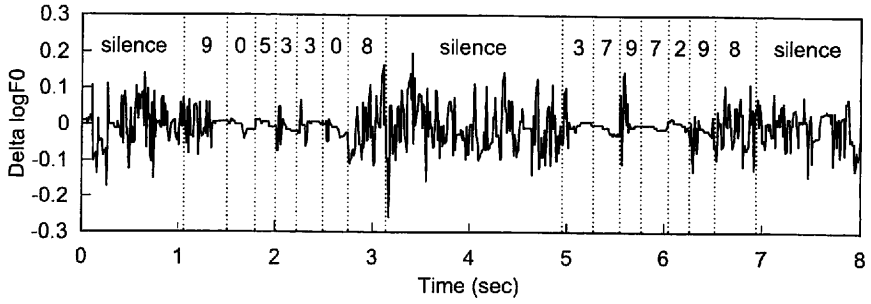
$$= \frac{d \log F_0}{dF_0} \cdot \frac{dF_0}{dt} \tag{5}$$

$$= \frac{1}{F_0} \cdot \Delta F_0 \tag{6}$$

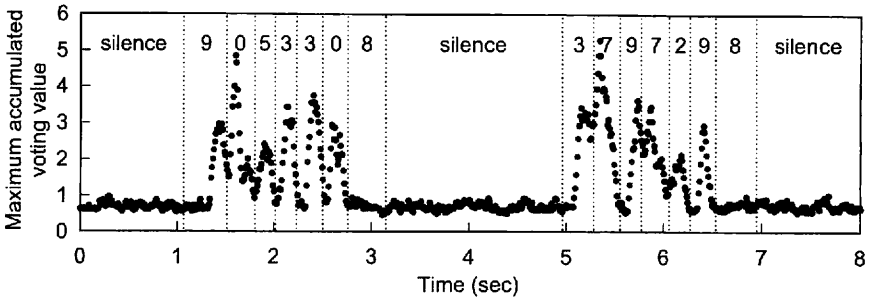
ΔF_0 is directly computed from the line extracted by the Hough transform. This $\Delta \log F_0$ value extracted by the Hough transform is denoted as "DLF0".

An example of the time functions of the DLF0 and MAVV is shown in Fig. 3. A male speaker's utterance, "9053308" "3797298", with white noise added at 20 dB SNR is shown. The dot lines indicate digit boundaries obtained by forced alignment using clean speech data and usual spectral-feature HMMs. Since DLF0 represents the F_0 transition, it should be useful for detecting digit boundaries. In unvoiced and pause periods, the DLF0 fluctuates more than in voiced periods. The MAVV in unvoiced and pause periods is much smaller than in voiced periods. These features are expected to be effective for detecting boundaries between voiced and unvoiced/pause periods.

The segmental and prosodic feature vectors are combined for each frame to build a segmental-prosodic feature vector.



(a) $\Delta \log F_0$ value extracted by the Hough transform (DLF0)



(b) Maximum accumulated voting value (MAVV)

Fig. 3 An example of the prosodic features in Japanese connected digit speech for a male speaker's utterance, "9053308" "3797298", with 20 dB SNR white noise.

Table 1 List of SP-HMMs (Segmental-Prosodic HMMs). SP-HMM is denoted by either "LC-SYL,PM" or "SYL+RC,PM". "LC-SYL" indicates the left-context dependent syllable and "SYL+RC" indicates the right-context dependent syllable. "PM" indicates F_0 pattern which is either rising ("U"), falling ("D"), or flat ("F").

digit	model		
0	ze+ro,U	ze+ro,D	ze+ro,F
/zero/	ze-ro,U	ze-ro,D	ze-ro,F
1	i+chi,U	i+chi,D	i+chi,F
/ichi/	i-chi,U	i-chi,D	i-chi,F
2	ni+i,U	ni+i,D	ni+i,F
/ni:/	ni-i,U	ni-i,D	ni-i,F
3	sa+N,U	sa+N,D	sa+N,F
/saN/	sa-N,U	sa-N,D	sa-N,F
4	yo+N,U	yo+N,D	yo+N,F
/yoN/	yo-N,U	yo-N,D	yo-N,F
5	go+o,U	go+o,D	go+o,F
/go:/	go-o,U	go-o,D	go-o,F
6	ro+ku,U	ro+ku,D	ro+ku,F
/roku/	ro-ku,U	ro-ku,D	ro-ku,F
7	na+na,U	na+na,D	na+na,F
/nana/	na-na,U	na-na,D	na-na,F
8	ha+chi,U	ha+chi,D	ha+chi,F
/hachi/	ha-chi,U	ha-chi,D	ha-chi,F
9	kyu+u,U	kyu+u,D	kyu+u,F
/kyu:/	kyu-u,U	kyu-u,D	kyu-u,F
	sil	sp	

3.3 Multi-Stream Syllable HMMs

3.3.1 Basic Structure of Syllable HMMs

Since CV syllable transition and the change of F_0 transitions, such as "rising", "falling" and "flat", are highly related, the segmental and prosodic information are integrated using syllabic unit HMMs. Our preliminary experiments show that syllable HMMs and tied-state triphone HMMs have approximately the same digit recognition accuracy for the connected digit task.

The integrated syllable HMM denoted by "SP-HMM (Segmental-Prosodic HMM)" is modeled by taking both the context and the F_0 transition into account. Table 1 is the list of SP-HMMs used in our experiments. Each Japanese digit uttered continuously with other digits can be modeled by a concatenation of two syllables (morae). Even "2" (/ni/) and "5" (/go/) can be modeled by two syllables since their final vowel is usually lengthened as /ni:/ and /go:/. The context of each syllable is considered only within each digit in our experiment. Therefore, the SP-HMM can be denoted by either a left-context dependent syllable "LC-SYL,PM" or a right-context dependent syllable "SYL+RC,PM", where "PM" indicates an F_0 transition pattern which is either rising (U), falling(D) or flat(F). For example, "the first syllable /i/ of "1" (/ichi/) which has rising F_0 transition" is denoted as "i+chi,U". Each SP-HMM has a standard left-to-right topology with $n \times 3$ states, where n is the number of phonemes in the syllable. The "sil" and "sp" models are used for a silence between digit strings and a short pause between digits, respectively.

3.3.2 Multi-Stream Modeling

SP-HMMs are modeled as multi-stream HMMs. In recognition, the probability $b_j(\boldsymbol{O}_{SP})$ of generating segmental-prosodic observation $\boldsymbol{O}_{SP}$ at state j is calculated by:

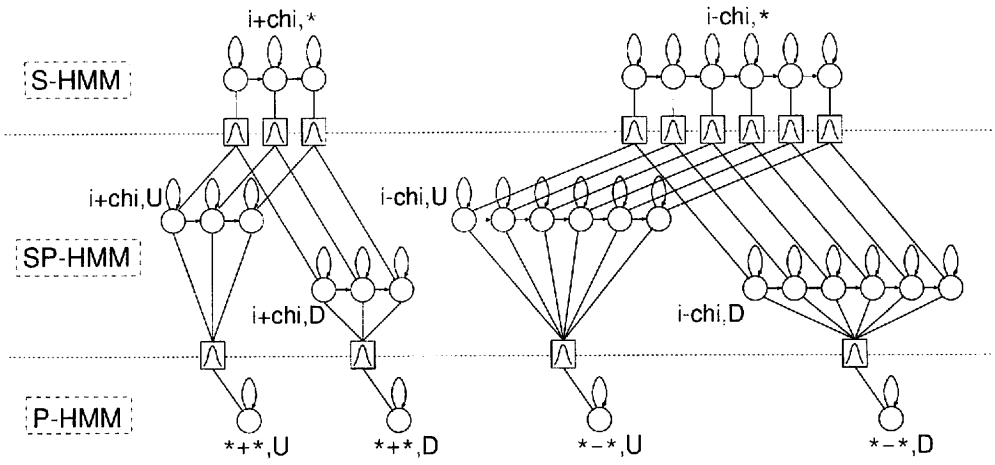


Fig. 4 Building SP-HMMs using a tied-mixture technique. S-HMMs and P-HMMs are trained using segmental features and prosodic features, respectively.

$$b_j(\boldsymbol{O}_{SP}) = b_j(\boldsymbol{O}_S)^{\lambda_S} \cdot b_j(\boldsymbol{O}_P)^{\lambda_P} \quad (7)$$

where $b_j(\boldsymbol{O}_S)$ is the probability of generating segmental feature vectors $\boldsymbol{O}_S$, and $b_j(\boldsymbol{O}_P)$ is the probability of generating prosodic feature vectors $\boldsymbol{O}_P$. λ_S and λ_P are weighting factors for the segmental stream and the prosodic stream, respectively. They are constrained by $\lambda_S + \lambda_P = 1$ $(0 \leq \lambda_S, \lambda_P \leq 1)$.

3.3.3 Building SP-HMMs

Syllable HMMs for segmental and prosodic feature vectors are separately made and combined to build SP-HMMs using a tied-mixture technique as follows:

1. "S-HMMs (Segmental HMMs)" are trained by segmental features only. They can be denoted by either "`LC-SYL,*`" or "`SYL+RC,*`". Here, "*" (wild card) means that HMMs are built without considering the F_0 transitions, "U", "D" and "F". The total number of S-HMM states is the same as the number of SP-HMM states. Twenty S-HMMs including "`sil`", "`sp`" are trained.
2. Training utterances are segmented into syllables by the forced-alignment technique using the S-HMMs, and one of the F_0 transition labels, "U", "D" or "F", is manually given to each segment according to the actual F_0 pattern.
3. "P-HMMs (Prosodic HMMs)" are trained by prosodic feature vectors within these segments, according to the F_0 transition label. Eight separate models, "`*-*,U`", "`*+*,U`", "`*-*,D`", "`*+*,D`", "`*-*,F`", "`*+*,F`", "`sil`" and "`sp`", are made. Each P-HMM has a single state, since it has been found that syllabic F_0 contours in Japanese can be approximated by a line function [9] and that the DLF0 can be expected to be almost constant in each syllable.
4. The S-HMMs and P-HMMs are combined to make SP-HMMs. Gaussian mixtures in the segmental stream of SP-HMMs are tied with corresponding S-HMM mixtures, while the mixtures in the prosodic stream are tied with corresponding P-HMM mixtures. Figure 4 shows the integration process. In this example, the mixtures of SP-HMM "`i+chi,U`" are tied with S-HMM "`i+chi,*`" and P-HMM "`*+*,U`".

3.4 Dictionary and Grammar

In the recognition dictionary, each digit has three variations according to the F_0 transitions. For instance, variations of "1" are "`i+chi,U i-chi,U sp`", "`i+chi,D i-chi,D sp`", and "`i+chi,F i-chi,F sp`". This means that the F_0 transition pattern does not change within the period of each digit. The recognition grammar is created so that all digits can be connected with no restrictions.

4. Experiments

4.1 Database

A speech database was collected from 11 male speakers in a clean/quiet condition. The database consists of utterances of 2-8 connected digits with an average of 5 digits. Each speaker uttered the digit strings, separating each string with a silence period. 210 connected digits and approximately 229 silence periods were collected per speaker.

Experiments were conducted using the leave-one-out method: data from one speaker were used for testing while data from all other speakers were used for training, and this process was rotated for all speakers. Training data were clean utterances, and testing data were contaminated with either white, in-car, exhibition-hall, or elevator-hall noise provided from JEIDA [10] at three SNR levels: 5, 10 and 20 dB. Accordingly, 11 speaker-independent experiments were conducted, and a mean word accuracy was calculated as the measure of recognition performance.

4.2 Evaluation of F_0 Extraction

The noise robustness of the F_0 extraction method using the Hough transform was evaluated. The F_0 extraction error (Hz) was defined as the RMS error between F_0 values extracted from clean utterances and that from noise-added utterances. Since F_0 cannot be observed in voiceless sounds, the errors were measured using the following digits having only voiced frames: /zero/, /ni:/, /yoN/, /go:/, and /nana/.

Table 2 shows the F_0 extraction errors when using the extraction method with or without the Hough transform at various SNR conditions. Errors for four kinds of noises are averaged at each SNR condition. Using the Hough transform greatly improves the robustness of F_0 extraction: extraction errors become roughly half by using the Hough transform in all SNR conditions.

4.3 Digit Recognition Results

Training and testing were performed using HTK [11]. In our preliminary experiments, the best S-HMM recognition performance (baseline) was obtained when the number of mixtures in each S-HMM was four. Therefore, we conducted experiments for selecting the best number of mixtures in the prosodic stream (P-HMMs) in SP-HMMs tied to the four mixture S-HMMs. The best performance of SP-HMMs was obtained when four mixture P-HMMs were used for all experiments.

4.3.1 Comparison of Various Prosodic Feature Vectors

We first investigated recognition performance in various conditions of prosodic feature vectors, in order to determine the effects of the prosodic features (DLF0 and MAVV) extracted by the Hough transform. Table 3 shows three kinds of prosodic feature vectors, **H-D**, **H-M**, and **H-DM**, built using the DLF0 and MAVV. For confirming the effect of the Hough transform on recognition results, two prosodic features were prepared without using the Hough transform for comparison; one was the $\Delta \log F_0$ value, which was computed by linear smoothing of $\log F_0$ values using 90 ms windows and is denoted as "DLF0' ". The other was a cepstral peak value, denoted as "CPV", which was used in place of MAVV. The table also shows three additional kinds of prosodic feature vectors, **NH-D**, **NH-C**, and **NH-DC**, which were composed by combination of the DLF0' and CPV. Consequently, six kinds of experiments were conducted according to the variation of prosodic feature vectors.

Figure 5 shows the digit error rates obtained by the SP-HMMs using various prosodic feature vectors at each SNR condition. The baseline performance by S-HMMs is also shown in the figure for comparison. The error rates for four kinds of noises are averaged at 20, 10, and 5 dB SNR, respectively. The segmental and prosodic stream weights and insertion penalties were optimized for each noise condition. Since all SP-HMMs reduce error rates except when using **NH-D**, the proposed modeling is effective for increasing noise robustness. The prosodic feature vectors extracted by the Hough transform, **H-D**, **H-M**, and **H-DM**, yield better performance than **NH-D**, **NH-C**, and **NH-DC**, respectively. This means that the Hough transform is advantageous for improving recognition performance. Among the feature vectors extracted, **H-DM** achieves the best performance. It is also found that the effects of the DLF0 and MAVV are additive.

As a supplementary experiment, we compared SP-HMMs with S-HMMs for digit boundary detection capability under noisy environments. Noise-added utterances

Table 2 F_0 extraction errors (Hz) by the extraction method with or without the Hough transform at various SNR conditions.

SNR	w/o Hough transform	with Hough transform
20 dB	11.4	5.8
10 dB	16.7	8.6
5 dB	20.5	11.2

Table 3 Six kinds of prosodic feature vectors.

Prosodic feature vector	Using Hough transform	Vector component (dim.)
H-D	yes	DLF0 (1)
H-M	yes	MAVV (1)
H-DM	yes	DLF0, MAVV (2)
NH-D	no	DLF0' (1)
NH-C	no	CPV (1)
NH-DC	no	DLF0', CPV (2)

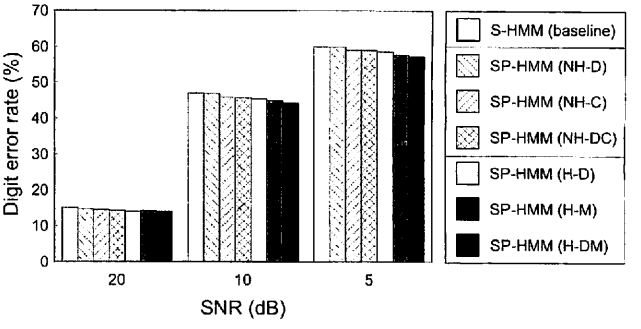


Fig. 5 Comparison of the digit error rates in various prosodic feature vectors.

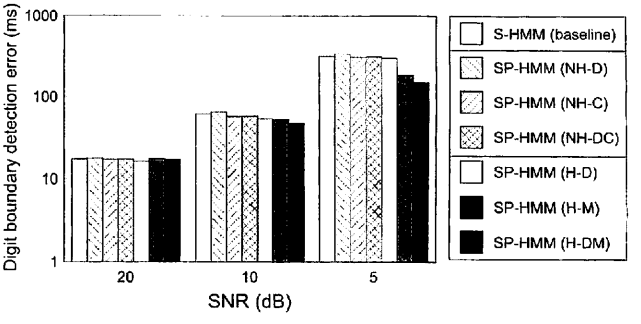


Fig. 6 Comparison of the digit boundary detection errors in various prosodic feature vectors.

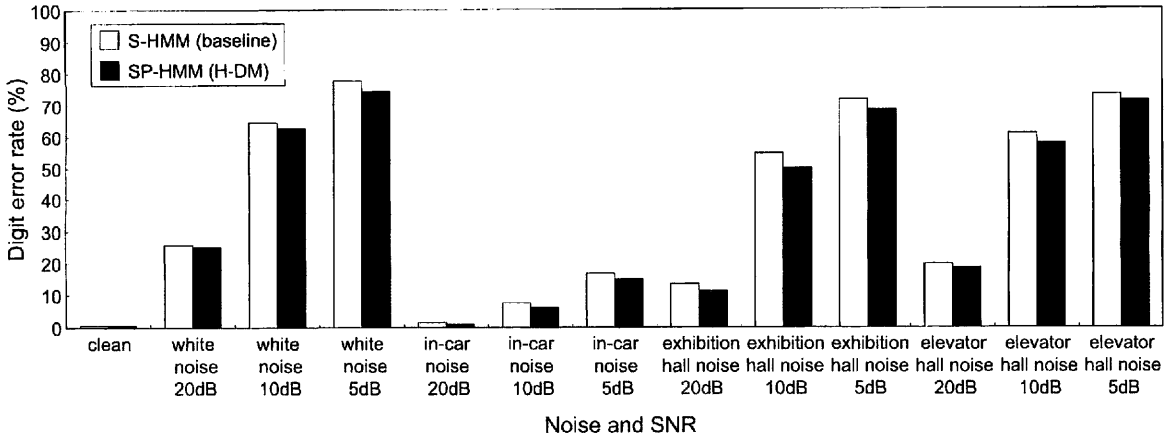


Fig. 7 Comparison of the digit error rates by SP-HMMs and S-HMMs in various noise and SNR conditions.

and clean utterances were segmented by these models using the forced-alignment technique. The digit boundary detection errors (ms) were computed by comparing the detected boundary locations in noise-added utterances with that in clean utterances. Figure 6 shows the boundary detection errors by S-HMMs and SP-HMMs when using various prosodic feature vectors. The errors for four kinds of noises were also averaged at each SNR condition. This figure displays a similar tendency as Figure 5; 1) SP-HMMs using **NH-C** and **NH-DC** slightly reduce error rates, 2) using **H-D**, **H-M**, and **H-DM** extracted by the Hough transform significantly reduces the errors, and 3) the best results are obtained when using **H-DM**. Consequently, we attribute the better recognition performance to precise boundary detection. The mean digit boundary detection error rate was reduced by 23.3% for 10 dB SNR utterances and 52.2% for 5 dB SNR utterances by using the SP-HMMs with **H-DM**.

In the following experiments, **H-DM** is used as the prosodic feature vector in SP-HMMs.

4.3.2 Results in Various Noise Conditions

The digit error rates using SP-HMMs with **H-DM** in various noise and SNR conditions are shown in Fig. 7. Recognition performance was improved in all kinds of noise conditions. The best improvement using SP-HMMs was observed when exhibition-hall noise was added at 10 dB SNR; digit accuracy was improved by 4.5% from 45.3% to 49.8% by the prosodic information.

4.3.3 Results for Each Speaker

In Fig. 8, the digit error rates by S-HMMs and SP-HMMs are shown for each speaker. In this experiment, 10 dB exhibition-hall noise was added to the test set. The improvement was observed for every speaker, which means that the proposed method is useful for speaker-independent recognition.

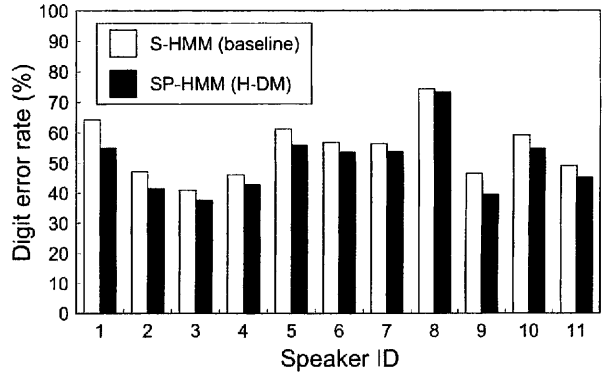


Fig. 8 Comparison of the digit error rates by SP-HMMs and S-HMMs for each speaker.

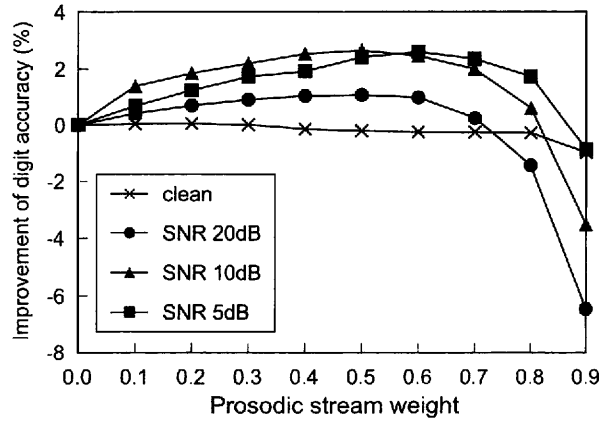


Fig. 9 Improvement of digit accuracy as a function of prosodic stream weight (λ_P) in each SNR condition.

4.3.4 Effects of the Prosodic Stream Weight

Figure 9 shows the improvement of digit recognition accuracy as a function of the prosodic stream weight λ_P at each SNR. Results for four kinds of noises are averaged at 20,

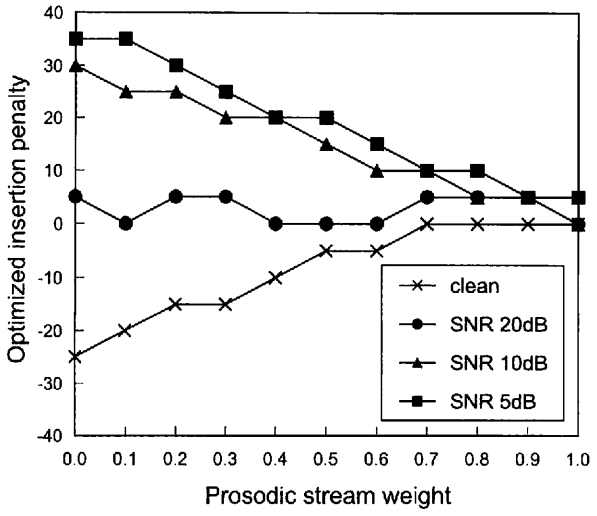


Fig. 10 Optimized insertion penalty as a function of prosodic stream weight (λ_P) in white noise condition.

10, and 5 dB SNR, respectively. The improvement using the SP-HMMs was observed over a wide range: $0.0 < \lambda_P \leq 0.7$ in all the noise conditions. Best results were obtained when λ_P was set around 0.6, irrespective of the SNR level.

Figure 10 shows the optimum insertion penalty as a function of the prosodic stream weight λ_P in the white noise condition. In noisy conditions, if the prosodic stream weight is low, we need to set the insertion penalty high to compensate for the low reliability of segmental features. Since prosodic features are effective for digit boundary detection, the higher the prosodic stream weight becomes, the lower the optimum insertion penalty becomes. Similar results were obtained for other noise conditions. The control range of the optimum insertion penalties in the best prosodic stream weight condition ($\lambda_P = 0.6$) is approximately a half of the range for the condition without using the prosodic information. This means that the prosodic features are effective for reducing the difficulty of adjusting the insertion penalty.

5. Conclusions

This paper proposed a recognition method using multi-stream syllable HMMs combining segmental and prosodic information. In order to robustly extract the prosodic features, we proposed an F_0 extraction method using the Hough transform. It was confirmed that the proposed method is robust in various noise conditions. In the experimental results of connected digit speech recognition, improvements were observed for every speaker and over a wide range of prosodic stream weights. It was also found that, by using the prosodic features, sensitivity of the insertion penalty in the decoder can be reduced.

Our future works include: 1) investigation of the SP-HMM topology, 2) study on combination with the adaptation method such as the MLLR technique, and 3) evaluation by more general recognition tasks. For future work on 3), a large speech corpus with prosodic labels is needed.

References

[1] Y. Sagisaka, N. Campbell, and N. Higuchi, eds., Computing Prosody, Part IV, Springer-Verlag, New York, 1997.

[2] G.Y. Chung and S. Seneff, "A hierarchical duration model for speech recognition based on the ANGIE framework," Speech Commun., vol.27, no.2, pp.113–134, March 1999.

[3] K. Iwano and K. Hirose, "Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition," Proc. Int. Conf. on Acoustics, Speech and Signal Process., vol.1, pp.133–136, Phoenix, AZ, March 1999.

[4] K. Hirose, N. Minematsu, and M. Terao, "Statistical language modeling with prosodic boundaries and its use for continuous speech recognition," Proc. Int. Conf. on Spoken Language, vol.2, pp.937–940, Denver, CO, Sept. 2002.

[5] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," Proc. European Conf. Speech Communication and Technology, vol.1, pp.311–314, Budapest, Sept. 1999.

[6] T. Shinozaki and S. Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," Proc. Int. Conf. on Acoustics, Speech and Signal Process., vol.1, pp.729–732, Orlando, FL, May 2002.

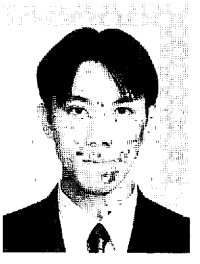
[7] S. Kori, "Onsei no tokucho kara mita bun," Nihongogaku, vol.15, no.8, pp.60–70, 1996.

[8] P.V.C. Hough, "Method and means for recognizing complex patterns," U.S. Patent #3069654, 1962.

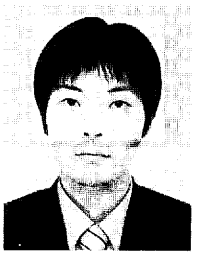
[9] K. Hirose and K. Iwano, "A method of representing fundamental frequency contours of Japanese using statistical models moraic transition," Proc. European Conf. Speech Communication and Technology, vol.1, pp.311–314, Rhodes, Greece, Sept. 1997.

[10] http://www.sunrisemusic.co.jp/dataBase/fl/noisedata01_fl.html

[11] http://htk.eng.cam.ac.uk/



Koji Iwano received a B. E. degree in information and communication engineering in 1995, and a M. E. and Ph.D. degrees in information engineering respectively in 1997 and 2000 from the University of Tokyo. He is currently an Assistant Professor at Tokyo Institute of Technology, Department of Computer Science. He is a member of the International Speech Communication Association (ISCA), the Information Processing Society of Japan (IPSJ), and the Acoustical Society of Japan (ASJ).



Takahiro Seki received a B. E. degree in information engineering in 2000, and a M. E. degree in information science and engineering in 2002 from the Tokyo Institute of Technology. He is currently with IBM Global Service-Japan.

Sadaoki Furui is currently a Professor at Tokyo Institute of Technology, Department of Computer Science. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 400 published articles. He is a Fellow of the IEEE and the Acoustical Society of America. He served as President of the Acoustical Society of Japan (ASJ) from 2001 to 2003 and he is now President of the International Speech Communication Association (ISCA) and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He is on the Board of Governors of the IEEE Signal Processing Society. He has served as Editor-in-Chief of the Transaction of the IEICE and an Editor-in-Chief of Speech Communication. He has received the Yonezawa Prize, the Paper Award and the Achievement Award from the IEICE (1975, 88, 93, 2003, 2003), and the Sato Paper Award from the ASJ (1985, 87). He has received the Senior Award from the IEEE ASSP Society (1989) and the Achievement Award from the Minister of Science and Technology, Japan (1989). He has received the Book Award from the IEICE (1990). In 1993, he served as an IEEE SPS Distinguished Lecturer.