HOKKAIDO UNIVERSITY

Title	Noise robust speech recognition using spectral subtraction and F0 information extracted by Hough transform
Author(s)	Yasui, Hideki; Shinoda, Koichi; Furui, Sadaoki; Iwano, Koji
Citation	Proceedings : APSIPA ASC 2009 : Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, 631-634
Issue Date	2009-10-04
Doc URL	http://hdl.handle.net/2115/39775
Type	proceedings
Note	APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. 4-7 October 2009. Sapporo, Japan. Poster session: Automatic Speech Recognition (6 October 2009).
File Information	TP-P1-4.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Noise robust speech recognition using spectral subtraction and F_0 information extracted by Hough transform

Hideki Yasui*, Koichi Shinoda*, Sadaoki Furui*, and Koji Iwano†
* Tokyo Institute of Technology, Meguro-ku, Tokyo, 152-8552, Japan
E-mail: {yasui,shinoda,furui}@cs.titech.ac.jp
† Tokyo City University, Tsuzuki-ku, Yokohama, 224-8551, Japan
E-mail: iwano@tcu.ac.jp

***Abstract*—We propose a noise robust speech recognition method based on combining novel features extracted from fundamental frequency (F_0) information and spectral subtraction. F_0 features have been shown to be effective in speech recognition in noisy environments. Recently, F_0 features obtained by Hough transform were developed for concatenated digit recognition and significantly improved recognition performance of noisy speech. This paper proposes novel features based on Hough transform for large-vocabulary continuous speech recognition. In addition, spectral subtraction is applied before Hough transform to remove static noise. The proposed method was tested using the Japanese Newspaper Article Sentences (JNAS) database. Word accuracy was improved in all noise conditions, with the best absolute improvement being 2.6 points in percentage when station noise was added at 10 dB SNR.**

I. Introduction

Considerable progress is being made in the field of large-vocabulary continuous speech recognition. When there is no surrounding noise, speech can be recognized with high accuracy, but speech recognition technology has not yet adequately overcome the problem of noisy environments.

Numerous feature extraction methods have been proposed to achieve robustness against noise (e.g.[1],[2]). Several studies have shown that fundamental frequency (F_0) information is useful for enhancing speech recognition performance in a noisy environment (e.g.[3],[4]). Since F_0 contours represent phrase intonation and word accent patterns, they can be expected to be useful in detecting prosodic phrases and word boundaries. In a recent study, F_0 features extracted by Hough transform were applied to connected digit recognition in noisy environments [5]. In this method, noise robust F_0 features were extracted by considering the time continuity of the F_0 pattern. Models were trained by manually assigning the F_0 transition labels "Up", "Down", or "Flat" to each segment. This method was reported to achieve a significant improvement in recognition in various noisy environments: white, in-car, exhibition-hall, or elevator-hall noise databases. In large-vocabulary continuous speech recognition (LVCSR) tasks, however, F_0 transition labels are not always assigned. In addition, the computational costs for Hough transform need to be reduced for real applications.

This paper proposes a robust speech recognition method using spectral subtraction [6] and F_0 information extracted by Hough transform for LVCSR. The method includes two novel F_0 features that are effective even when F_0 transition labels are not available. In addition, we used spectral subtraction as a preprocess for feature extraction. Since the effect of the F_0 features does not change after static noise is removed, more improvement could be expected to be provided by combining spectral subtraction and F_0 information.

II. Proposed method

We propose a noise robust speech recognition method using spectral subtraction and F_0 information extracted by Hough transform. A flow chart of the proposed method is shown in Fig. 1.

A. Spectral subtraction

Spectral subtraction has been shown to be effective for steady noise and requires relatively small computational costs. It is performed using the following equation:

$$|\hat{S}(f)|^2 = \max\{|X(f)|^2 - \alpha|\hat{N}(f)|^2, \beta|\hat{N}(f)|^2\}, \quad (1)$$

where $|\hat{S}(f)|^2$ is the estimated power spectrum, $|X(f)|^2$ is the observed power spectrum, and $|\hat{N}(f)|^2$ is the noise power spectrum estimated from a non-speech segment. α and β are subtraction coefficients. In this paper, α was set to 1.0 and β to 0.24.

B. F_0 information extraction

1) F_0 information extraction by Hough transform: A noise robust F_0 information extraction method using Hough transform was recently proposed [5]. In this method, time-cepstrum images were cut out using a suitable window size, and then the most dominant straight line was taken out of each time-cepstrum image using Hough transform [7]. This method effectively utilizes the time-continuous nature of F_0 information under noisy conditions, as we explain in detail below.

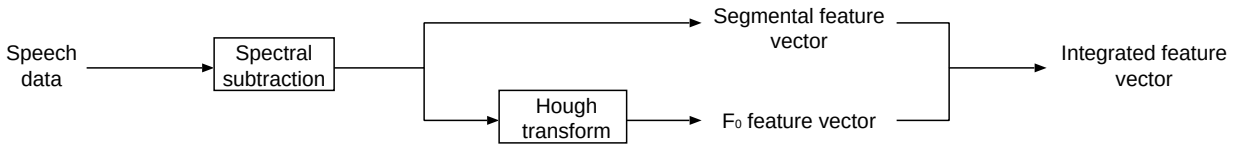

Fig. 1. A flow chart of the proposed method.

First, speech waveforms were sampled at 16 kHz and transformed to 256 dimensional cepstrum. Then, a 32 ms-long Hamming window was used to extract frames every 10 ms. Since the peak value tends to be large in a low-dimensional cepstrum under noise, the following weight factor k_d was multiplied to the cepstrum coefficients of d-th dimensional in the low dimension (30-140 dimension) to reduce its effect:

$$k_d = 0.6 + 0.4 \sin\left(\frac{d-30}{140-30} \times \frac{\pi}{2}\right). \quad (2)$$

Next, a nine-frame moving window was applied at every frame interval to extract a time-cepstrum image for line information extraction by Hough transform.

Suppose the time-cepstrum image consists of n pixels (x_i, y_i) $(i = 1, \cdots, n)$. In Hough transform, every pixel on the x-y plane is transformed to a line on the m-c plane as

$$c = -x_i m + y_i \ (i = 1, \cdots, n). \quad (3)$$

The brightness value of each pixel on the x-y plane is accumulated at every point on the line. This process is called "voting" to the m-c plane. Only points that have values larger than a predetermined threshold on the x-y plane are used for voting to reduce the computational costs. The threshold was set to be 0.1 from the results of our preliminary experiment. After voting, the peak point $(\hat{m}, \hat{c})$ with the maximum accumulated voting value on the m-c plane is transformed to a line on the x-y plane by the following equation:

$$y = \hat{m}x + \hat{c}. \quad (4)$$

An F_0 value(Hz) is obtained from the cepstrum index at the center point of the detected line.

2) Type of F_0 information: Iwano *et al.* [5] proposed two F_0 features, F and V:

F: $\Delta \log F_0 \approx \Delta F_0 / F_0$, where ΔF_0 is directly computed from the line extracted by the Hough transform.

V: The accumulated voting value obtained in the Hough transform at the peak point $(\hat{m}, \hat{c})$. It represents the degree of temporal continuity in the F_0.

In their method, the F_0 transition labels "Up", "Down", and "Flat" for each digit were utilized to recognize concatenated digits. In LVCSR, however, it is not feasible to prepare such labels for every word in the dictionary. Therefore, F_0 feature "F" proposed by Iwano *et al.* [5] might not be effective. To overcome this problem, we proposed two novel features, D and N:

D: Dynamic feature of the voting value V. D_t, which is the F_0 feature D in the t-th frame, is computed as follows:

$$D_t = \frac{\sum_{i=1}^{2} i(V_{t+i} - V_{t-i})}{10}, \quad (5)$$

where V_t is V in the t-th frame.

N: Normalized V, which is computed as follows:

$$N = \frac{V - V_{\min}}{V_{\max} - V_{\min}}, \quad (6)$$

where $V_{\max}$ and $V_{\min}$ are the maximum and minimum values of V in each utterance, respectively.

C. Integration of Segmental and F_0 information

We integrated the conventional cepstrum features representing segmental information and the proposed F_0 features by using a multi-stream HMM scheme. In each state j, the probability $b_j(\vec{O}_{SF})$ of generating integration vector $\vec{O}_{SF}$ is calculated as:

$$b_j(\vec{O}_{SF}) = b_j(\vec{O}_S)^{\lambda_S} \cdot b_j(\vec{O}_F)^{\lambda_F} \ , \quad (7)$$

where $b_j(\vec{O}_S)$ is the probability of generating segmental feature vector $\vec{O}_S$ and $b_j(\vec{O}_F)$ is the probability of generating F_0 feature vector $\vec{O}_F$. λ_S and λ_F are weighting factors for the segmental and F_0 streams, respectively ($\lambda_S + \lambda_F = 1$).

To make multi-stream HMMs, segmental HMMs (S-HMMs) which are triphone models with 3000 states are first trained using the segmental features. Next, each training utterance is segmented into a sequence of phonemes by the forced-alignment technique using the S-HMMs. Then, a GMM for the F_0 features is trained for each state of each phoneme using the time labels given by the alignment. Finally, the phoneme S-HMM and state GMMs are combined to construct a multi-stream HMM for each phoneme.

III. Experiments

A. Experimental conditions

The Japanese Newspaper Article Sentences (JNAS) database was used for the evaluation of the proposed method. In this database, 260 (130 males and 130 females) speakers' utterances were used as training data and 46 (23 males and 23 females) speakers' utterances were used as test data. The number of utterances for each speaker was 100 on average. The original clean data were used for training HMMs. The test data were synthesized with station, elevator-hall, train, and

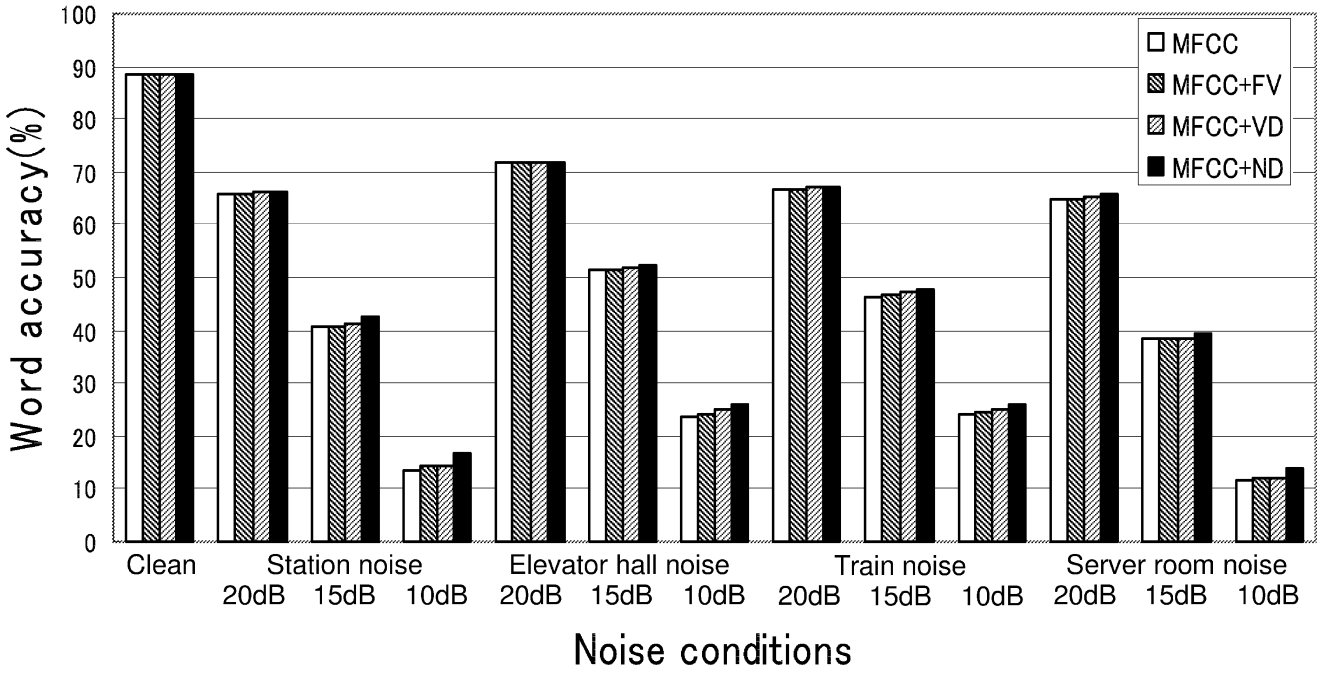

Fig. 2. Recognition accuracies for three F_0 features (without spectral subtraction). MFCC represents the results obtained by using segmental HMMs only. MFCC+FV, MFCC+VD, and MFCC+ND are the results obtained by multi-stream HMMs using F_0 features (F, V), (V, D), and (N, D) respectively.

server-room noise at three SNR levels: 10, 15 and 20 dB. We used a trigram language model and a dictionary consisting of 60,000 words.

Each segmental feature vector had 25 elements consisting of 12 MFCC, their delta, and the delta log energy. The window length was 25 ms and the frame interval was 10 ms. Cepstral mean normalization (CMN) was applied to each utterance. We investigated three kinds of F_0 feature vectors, (V, F), (V, D), and (N, D). Each feature vector was combined with the segmental feature vector in each frame, as described in Section 4. We set the range for voting as $-20 < m < 20$ and $-300 < c < 300$ in the m-c plane for Hough transform. The image size of the m-c plane was set to 200×1200.

We used a Julius decoder [8] in our recognition experiments. The language weights and insertion penalties were optimized for clean conditions. The segmental and F_0 stream weights were optimized for each noise condition in the test environment.

B. Results

1) Effects of F_0 features: First, we compared the recognition accuracies for the different F_0 features. Fig. 2 shows the word accuracy using S-HMMs and integration-HMMs in various noise conditions. Word accuracy was not improved in MFCC+FV; from 43.2% to 43.4% on average over all the noise conditions. This may have been mainly because the F_0 transition labels were not available. MFCC+VD was effective in all kinds of noise conditions (43.8% on average). MFCC+ND showed the best performance (44.7% on average), which indicates that the mismatch of the F_0 feature caused by noise was reduced by the normalization.

2) Effects of combining F_0 features with spectral subtraction: Next, we evaluated combining the F_0 features with spectral subtraction. Here, spectral subtraction was carried out as a preprocess for both the segmental and F_0 features. As F_0 features, we used (N, D) which provided the best performance of the three F_0 features. The proposed method was effective in all kinds of noise conditions (52.5% on the average). The best improvement of 2.6 points from 25.1% to 27.7% was observed in the condition when station noise was added at 10 dB SNR.

3) Reducing computational cost of Hough transform: Since the computational costs for Hough transform are relatively high, we assessed the reduction in computational costs that could be achieved by reducing the size of the Hough plane. Table I shows the relationship between the size of the Hough plane and computational time. We were able to reduce the computational time by 21.9%, but this was still long. A decrease of 0.15 points in recognition accuracy was observed in the condition when station noise was added at 15 dB SNR.

IV. Conclusion

We proposed a robust speech recognition method based on using spectral subtraction and F_0 information extracted by Hough transform. We also proposed two novel F_0 features that are effective for large-vocabulary continuous speech recognition. This method was shown to be robust under various noise conditions. Word accuracy improved from 43.2% to 52.5% on average when we used the proposed method and test data

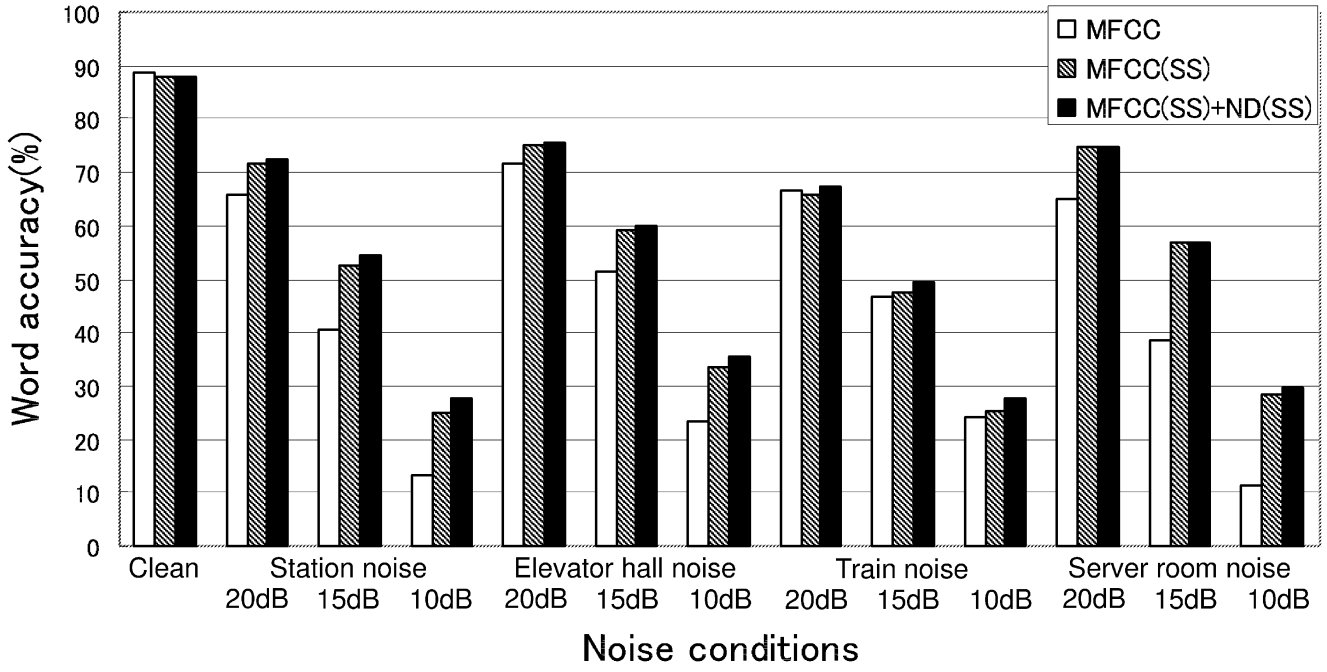

Fig. 3. Recognition accuracies of the proposed method using spectral subtraction and F_0 feature. MFCC represents the results obtained using segmental HMMs only, MFCC(SS) represents the results obtained using segmental HMMs with spectral subtraction, and MFCC(SS)+ND(SS) represents the results obtained by multi-stream HMMs using F_0 features (N, D) with spectral subtraction.

TABLE I

COMPUTATIONAL TIME WHEN F_0 FEATURES WERE EXTRACTED FROM SPEECH CEPSTRUM OF ONE UTTERANCE (12.6S). WE USED A COMPUTER WITH INTEL CORE 2 DUO 2.4 GHZ AND 2 G BYTE MEMORY.

Size of Hough plane	Computational time (s)
200×1200	192
100×600	42

synthesized with station, elevator-hall, train, and server-room noise at three SNR levels: 10, 15, and 20 dB. Future work will include efforts to further reduce the computational time required for Hough transform. We will also assess the effect of combining our multi-stream framework with noise robust techniques other than spectral subtraction.

REFERENCES

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans.SAP*, vol. 2, no. 4, pp. 578-589, 1994.

[2] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans.SAP*, vol. 4, no. 5, pp. 352-359, 1996.

[3] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment," *Proc. ICASSP2001*, pp. 125-128, 2001.

[4] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," *Proc. ICSLP2002*, pp. 1065-1068, 2002.

[5] K. Iwano, T. Seki, and S. Furui, "Noise robust speech recognition using F_0 contour extracted by Hough transform," *Proc. ICSLP2002*, pp. 941-944, 2002.

[6] S.F.Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans.ASSP*, vol. 27, no. 2, pp. 113-120, 1979.

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

[8] Julius (ver.3.4.2), http://julius.sourceforge.jp/en/julius.html.