3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ISOLATED WORD RECOGNITION USING MODELS FOR ACOUSTIC PHONETIC VARIABILITY BY LOMBARD EFFECT

Tadashi Suzuki, Kunio Nakajima and Yoshiharu Abe
Computer and Information Systems Laboratory
Mitsubishi Electric Corporation
5-1-1, Ofuna, Kamakura, Japan 247

Abstract

In noisy environment, performance of speech recognition system trained in quiet environment is degraded. We propose a new word recognition method using an acoustic phonetic variability model for Lombard effect that is one of the reasons for this degradation. In this method, difference between a spectral envelope of normal speech and that of Lombard speech is represented by the acoustic phonetic variability model, which are comprised of a non-linear warping function on spectral frequency domain for formant shift and spectral filters for changes of formant bandwidths and spectral tilt. Each model is trained with Lombard speech and provided for a sub-phoneme HMM. In Lombard speech recognition, the HMMs are modified with the acoustic-phonetic variability models, and the duration parameters are modified to compensate the word duration changes by Lombard effect. Recognition experiments without contamination-by-noise were conducted. The Lombard speech data was comprised of isolated 100 words spoken by 5 males hearing 90dB(SPL) pink noise through headphones. The recognition rate was 98.6% with this method, and 88.4% without the method.

1. Introduction

A speech recognition system, trained with speech data spoken in quiet environment, performs poorly in noise. The reason is not only a contamination of speech signal by noise, but also acoustic phonetic changes by Lombard effect that occur while speakers change their speech manner to communicate in so noisy environments as utterances are masked.

A method of transforming word templates of normal speech into Lombard speech using speaker adaptation technique has been proposed in [1]. In this method, based on a speech recognizer using a VQ codebook, each feature vector given by normal speech's codebook is replaced by the average of Lombard speech's feature vectors, corresponded to the codebook vector by DTW between normal and Lombard speech data.

The above method is applicable to continuous density HMMs of sub-phonemes by replacing each mean vector of normal speech HMMs with the average of Lombard speech feature vectors corresponded to the HMMs.

However, the recognition accuracy for Lombard speech with those HMMs is not satisfactory, since spectral modification by Lombard effect has variability depending on speaking effort or accent etc., even in stationary noise environment. Moreover change of word duration caused by Lombard effect degrades the recognition performance.

To cope with the problem, we propose a new word recognition method using an acoustic phonetic variability model by Lombard effect. It is expected that using the model we can represent the variability in degree of the modification by Lombard effect. In the model, the change in spectral envelope by Lombard effect is divided into three factors; formant shift, bandwidth reduction and spectral tilt change. For each factor, we define a function that has a parameter which controls degree of the spectral change.

Although the acoustic phonetic variability model is applicable to a variety of recognition methods from VQ-based one to continuous density HMMs, we describe a recognition method based on continuous density HMMs of sub-phonemes. Each sub-phoneme HMM, trained with normal speech, is provided with the acoustic phonetic variability model whose parameter is trained with the HMM and Lombard speech data. A mean vector of the HMM is transformed with its acoustic phonetic variability model to recognize Lombard speech. In addition, duration parameters of the HMM are modified according to duration changes observed in the Lombard speech data.

2. Acoustic phonetic variability model

The acoustic phonetic variability model is constructed with three functions for formant shift, bandwidth reduction and change of spectral tilt. The formant shift is represented by a dynamic frequency warping function obtained with DP matching between a spectral envelope of normal speech and spectral envelopes of Lombard speech. The bandwidth reduction is approximated by the weighted-sum of a peak-enhanced spectral envelope of normal speech and the original spectral envelope.

When the spectral envelope of normal speech is discretely represented by a sequence of log-powers as $\{S_i\}$, a modified spectral envelope by the acoustic phonetic variability model is given by

$$T_j = S_j + q_j(w) + \delta_j(w)$$

and

$$j = i + \theta_i(w) \qquad (i=0,1,\ldots,I) \qquad (1)$$

where $\{T_j\}$, a sequence of log-powers, represents the modified spectral envelope, and parameter i and j denote frequencies.

In equation (1), $q_j(w)$ is a log-power spectrum of a filter representing the bandwidth reduction, $\delta_j(w)$ is a log-power spectrum of a filter representing change of spectral tilt, and $\theta_i(w)$ is a frequency warping function representing formant shift. Parameter w is a variable which controls degree of the modifications. The two filters and one function are defined as

$$\theta_i(w) = w \cdot \Theta_i \qquad (2a)$$

$$q_i(w) = w \cdot Q_i \qquad (2b)$$

$$\delta_i(q) = w \cdot \Delta_i \qquad (2c)$$

where, Θ_i, Q_i, Δ_i denote the average modification of Lombard speech data. The Θ_i, Q_i, Δ_i are obtained with following procedures.

step 1. Obtain alignments of normal HMMs on Lombard speech data with a Viterbi algorithm.

step 2. Carry out following sub-procedures for each HMM.

2-1. Find the frequency warping function with which the least matching distortion is obtained averaging across all DP matchings between the spectral envelope given by the normal HMM and the corresponding spectral envelopes of Lombard speech.

2-2. Obtain mean difference between the spectral envelope given by the normal HMM and the spectral envelopes of Lombard speech, in which the influence of formant shift is removed by frequency-warping according to the warping function obtained in step 2-1.

2-3. Obtain g to minimize E given by equation (3),

$$E = \sum_i (D_i - g \cdot P_i)^2 \qquad (3)$$

where $\{D_i\}$ is the mean spectral difference obtained in step 2-2 and $\{P_i\}$ is a peak-enhanced spectral envelope derived from the spectral envelope given by the HMM. Obtained $\{g \cdot P_i\}$ represents a bandwidth reduction filter, and the remainder, $\{D_i - g \cdot P_i\}$, represents a spectral tilt modification filter.

step 3. Make modified HMMs from the normal HMMs, where each mean vector of the HMMs is transformed with the frequency warping function and two filters obtained in step 2.

step 4. Obtain new alignments with the modified HMMs on Lombard speech data, and return to step 2 to iterate the steps 2, 3 and 4 several times.

The parameter w can be set to a value equal or greater than 0. In case of the absence of Lombard modification, we can set the w to 0. In case of the existence of Lombard modification, we can obtain greater modification with larger value of w and the same modification as Lombard speech data by setting the w to 1.

The suitable value of w is thought to be varied according to vocal effort and accent. So it is desirable to determine the w to minimize the distortion between the modified HMM and each of the input feature vectors. To represent these variability in w, we provide plural values of w for every feature vectors.

[Duration compensation]

In the speech recognition method based on sub-phoneme HMMs, duration control technique is effective very much. But its performance is sensitive to duration change by Lombard effect.

We use a compensation method for the duration change using the alignments obtained while acoustic phonetic variability models are trained. In this method, we calculate the average ratio of normal duration to Lombard speech's duration for each sub-phoneme across the all speakers and multiply the normal speech duration by the ratio.

3. Experimental evaluation

3.1 Experimental conditions

Speech data used in our experiments were sets of isolated 100 words, the names of cities in Japan, spoken by 5 males. Lombard speech was produced by the speakers hearing pink noise through headphones, and thus noise-free.

The speech data were sampled at 10kHz, and pre-emphasized by a filter $(1 - 0.95 \cdot z^{-1})$. Using Hamming window (25.6msec), 12-order LPC analysis were carried out every 10msec, and 16 LPC cepstral coefficients were extracted.

The HMMs for normal speech were trained for each speaker using 4 sets of the 100 words spoken by the speaker hearing no noise. The acoustic-phonetic variability models were trained with 4 sets of 100 words spoken by the speaker hearing 90dB(SPL) pink noise. Another set was used for evaluation.

Every sub-phoneme HMMs had 1 state and 1 loop. A word model was represented by a string of sub-phoneme HMMs connected by null-transitions.

Frequency warping and filetering were carried out by matrix computation and addition in cepstrum domain. Thus peak-enhancement of a spectral envelope was realized by quefrency-weighting of the original cepstrum.

3.2 Evaluation by word recognition

In order to validate the acoustic phonetic variability model, we conducted recognition experiments using Lombard speech produced in 80 and 90dB(SPL) pink noises. Figure 1 shows the results, when the value of w was varied from 0 to 1.6 by 0.2. The filled symbols denote word error rates and blanked symbols denote average distortions. The squares denote the results for Lombard speech by 90dB(SPL) pink noise. This condition was as same as the training data. The circles denote the results for Lombard speech by 80dB(SPL).

For Lombard speech by 90dB(SPL), w=1.0 is optimum. For Lombard speech by 80dB(SPL), optimum value of w is 0.4 or 0.6. These results show that our model is able to represent the variability of Lombard modification.

3.3 Distribution of Lombard modification

Next, we investigated a distribution of optimum value of w for each feature vector of Lombard speech used in the training.

Figure 2 shows a histogram for the optimum values of w averaged for Lombard speech spoken by 5 males. As shown in the figure, it appears that the Lombard modifications of sub-phonemes have variability frame-by-frame.

3.4 Duration change

Duration parameters of sub-phoneme HMMs were comprised of mean values and standard deviations.

For the mean values and standard deviations, the ratios of normal speech to Lombard speech were calculated using alignments of HMMs on normal speech and Lombard speech. The Lombard speech was produced in 90dB(SPL) pink noise.

We observed that the mean durations of stationary parts of vowels and transient parts of voewls adjacent to word-tails were distinguishably extended for all speakers. Those facts agree with results in [2]. We also observed that for most of sub-phonemes, the standard deviations were larger than those of normal speech for all speakers.

3.5 Lombard speech recognition

Speaker dependent word recognition experiments were executed with noise-free Lombard speech by 90dB(SPL). The set of parameter Ws was set as {0.0, 0.8, 1.0, 1.2, 1.4} according to the result in 3.3. We selected the optimum W from the set frame-by-frame. Each duration parameter of sub-phoneme HMMs was modified with the average of all speakers' ratios obtained in 3.4.

Figure 3 shows the average word error rates for 5 speakers. In the figure, " normal " denotes recognition with normal HMMs, " adapt " denotes recognition with Lombard adapted HMMs, mean vectors of which are made by a speaker adaptation method. The error rate of 1.4%, the best performance, is achieved by our method " APVM "

4. Conclusion

In this paper, we have proposed a word recognition method using the acoustic phonetic variability model representing spectral changes of utterances by Lombard effect. The model is comprised of a non-linear warping function on spectral frequency domain for representing formant shift, and two spectral filters, one of which represents bandwidth reduction of formants, and the other represents mainly spectral tilt changes. A compensation method for changes in sub-phoneme duration has also been described.

Experimental evaluations were executed in speaker-dependent isolated word recognition based on continuous density HMMs of sub-phonemes. Simulated Lombard speech of Japanese 100 city names were spoken by 5 males hearing 90dB(SPL) pink noise through headphones.

From the experiments, in which the proposed method achieved better performance than the speaker adaptation method, the effectiveness of the proposed method has been confirmed.

References

[1] David B. Roe, " Speech Recognition with a Noise-Adapting Codebook, " Proc. ICASSP, pp. 1139-1142, 1987.

[2] B.J.Stanton, L.H.Jamieson, G.D.Allen, " Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions, " Proc. ICASSP, pp. 331-334, 1988.

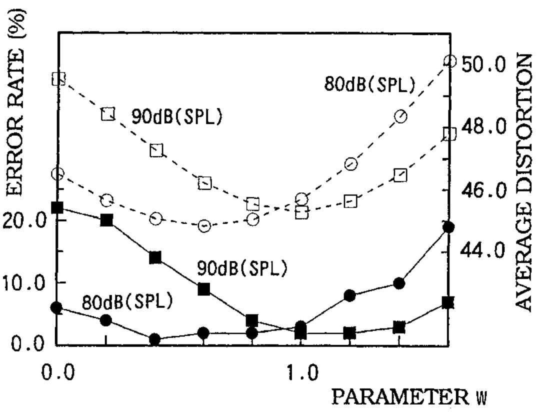

Figure 1: An example of word recognition performance changing parameter W.

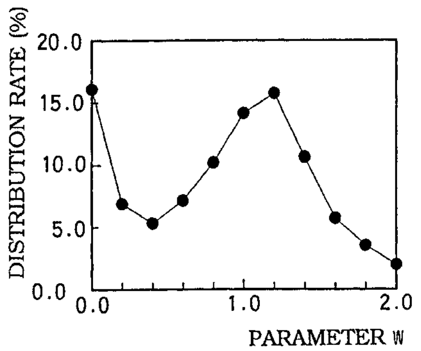

Figure 2: Distribution of optimum Ws of every feature vectors in Lombard speech data by 90dB(SPL) pink noise.

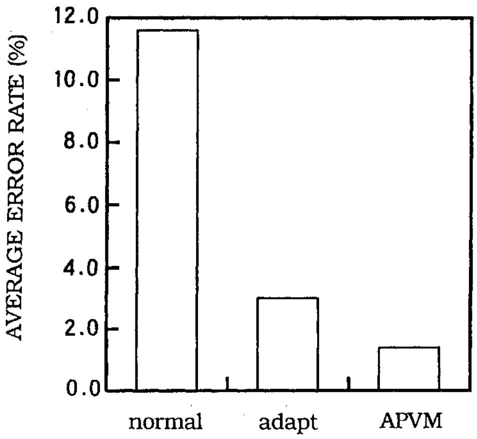

Figure 3: Results of Lombard speech recognition.