Robust Recognition of Fast Speech

SUMMARY This letter describes a robust speech recognition system for recognizing fast speech by stretching the length of the utterance in the cepstrum domain. The degree of stretching for an utterance is determined by its rate of speech (ROS), which is based on a maximum likelihood (ML) criterion. The proposed method was evaluated on 10-digits mobile phone numbers. The results of the simulation show that the overall error rate was reduced by 17.8\% when the proposed method was employed.

key words: robust speech recognition, rate of speech, maximum likelihood estimation

1. Introduction

The Rate of Speech (ROS) is one of the factors that affect the performance of automatic speech recognition (ASR). This is influenced by acoustical and the phonological differences associated with fast speech [1]–[4]. In fact, it has been shown that the variation in rate during the course of a conversation averages about 31\% and can be as high as 60\% [5]. This means that ROS-adaptive recognition would be highly desirable, in terms of improving the performance of recognition systems on conversational speech recognition tasks.

To implement an ROS-adaptive recognition system, two factors should be considered: how to measure speaking rate and how to modify incoming signals or models (e.g. HMM). In [2], the ROS was estimated by the inverse of the mean duration or the mean of the rates, where a hypothesized phonetic transcription from the underlying ASR is used. In [5], an artificial neural network (ANN) is used to estimate the ROS. The first spectral moment of the wideband energy envelope has also been used for estimating ROS [3].

Modification of speech/model according to the ROS of a given speech signal has not been sufficiently discussed to date. In [2], HMMs are modified by increasing the phone exit probability. Richardson et al. [7] proposed cepstrum length normalization (CLN), where the phone duration is normalized by stretching the length of the utterance in the cepstrum domain. In this method, however, the stretching factors are not optimally determined in the sense of maximizing recognition accuracy (or in the sense of maximum likelihood). Using an independent set of HMMs constructed from the fast speech corpus can also be used as an alternative way to cope with fast speech. This method, however, requires additional memory space for storing an independent set of HMMs.

In the proposed method, incoming utterances are classified by their ROS, the cepstrum coefficients are then stretched by corresponding stretching factor. A joint optimization method is proposed here, where the optimum ROS thresholds for classification and the optimum stretching factors are simultaneously determined in the sense of maximum likelihood criterion. The proposed algorithm finds classification thresholds and the scaling factors that maximize the likelihood between time scaled cepstrum coefficients and HMMs obtained from normal rate speech signals.

2. Rate-Adaptive ASR

A block diagram of the proposed ASR system is depicted in Fig. 1. The ROS of the incoming utterance is first estimated using the energy distribution obtained from a Mel-scale filter bank. With a given ROS, it is then determined how much a sequence of cepstral coefficients is stretched. Each part of the proposed ASR is described in more detail in the following subsections.

2.1 ROS Estimation

In this work, vowel detection algorithm [6] is employed to estimate the ROS of the incoming utterance. In this method, the prominent local maxima of the smoothed modified loudness function [6] are detected as vowel locations, the ROS is then computed by dividing the total number of peaks by the sentence duration. From our experimental results, it was observed that a method for estimating ROS by vowel detection produced the highest correlation (\( \rho = 0.79 \)) with the real ROS derived from the manually labelled utterances, among the other ROS estimation methods.

2.2 Determining an Optimal Classification Rule and a Set of the Stretching Factors

An important issue associated with a rate-adaptive ASR system is the determination of the optimal classification boundaries for ROS and a set of optimum stretching factors for each class. These two parameters should be determined in order to achieve maximum recognition accuracy. To this end, we first define a following log likelihood function involved with a set of the classification boundaries for ROS.
\[ \Theta = \{ \theta_0, \theta_1, \ldots, \theta_M \} \] and a set of the stretching factors \( \Gamma = \{ \gamma_1, \gamma_1, \ldots, \gamma_M \} \).

\[ \mathcal{L}(X, \Lambda, \Theta, \Gamma) = \sum_{m=1}^{M} \sum_{x \in S_m} \log p(f(x, \gamma_m)|\Lambda). \tag{1} \]

where \( \Lambda \) denotes an HMM parameter set and \( f(x, \gamma_m) \) is the modified cepstrum vector obtained by stretching the original cepstrum vector with a stretching factor \( \gamma_m \). Note that both \( \Theta \) and \( \Gamma \) are ordered set and \( \theta_0 = -\infty, \theta_M = +\infty \), where \( M \) is the number of classes for ROS. In (1), \( S_m \) is a set of the cepstrum vectors those have ROS belonging to class \( m \), i.e.

\[ S_m = \{ x \in X_s, \theta_m-1 \leq r(s) \leq \theta_m \}, \tag{2} \]

where \( r(s) \) is the ROS of the sentence \( s \) and \( X_s \) is the set of the cepstrum vectors derived from the sentence \( s \). Now, an optimal set of classification boundaries and a corresponding set of the stretching factors for a given HMM parameter set \( \Lambda \) are given by

\[ \{ \Theta^*, \Gamma^* \} = \arg \max_{\Theta, \Gamma} \mathcal{L}(X, \Lambda, \Theta, \Gamma). \tag{3} \]

To find \( \Theta^* \) and \( \Gamma^* \), a joint optimization algorithm is proposed. The proposed algorithm iteratively finds the optimum \( \Theta \) and \( \Gamma \). The overall procedure is as follows:

**Step-0. Initialization:** Given training set \( X \), an HMM set \( \Lambda \) is built using the Baum-Welch method. An initial \( \Gamma^{(0)} = \{ \gamma_1^{(0)}, \gamma_2^{(0)}, \ldots, \gamma_M^{(0)} \} \) is also built using an adequate method. Set threshold \( \epsilon, \Lambda^{(0)} = -\infty \) and \( i = 0 \).

**Step-1. Determining ROS classification boundaries:** With a previously determined \( \Gamma^{(i)} \), find the set of the classification boundaries that yields the maximum overall likelihood.

\[ \Theta^{(i)} = \arg \max_{\Theta} \mathcal{L}(X, \Lambda, \Theta, \Gamma^{(i)}). \tag{4} \]

This maximization problem can be formulated as finding the maximum likelihood path in the trellis where each node \( (n,m) \) in the trellis corresponds to the local likelihood for the \( n \)-th ROS value \( r_n \) and the \( m \)-th stretching factor \( \gamma_m \) as follows

\[ l(n,m) = \sum_{x \in S_n} \log p(f(x, \gamma_m)|\Lambda). \tag{5} \]

where \( S_n = \{ x \mid x \in X_s, r(s) = r_n \} \), i.e. a set of the cepstrum vectors derived from the sentence having the ROS value of \( r_n \). Hence, \( l(n,m) \) means that the sum of the log likelihoods of the modified cepstrum vectors by stretching with the \( m \)-th stretching factor \( \gamma_m \), where underlying cepstrum vectors are derived from the sentence having the ROS value of \( r_n \). A graphical explanation of this principle is shown in Fig. 2.

Assuming that \( r_n < r_{n+1}, \ 1 \leq n \leq N-1 \) and \( \gamma_m < \gamma_{m+1}, \ 1 \leq m \leq M-1 \), the maximum likelihood path can be
found by Viterbi decoding as follows

forward recursion:
\[
L(n, m) = \max_{k=m,m-1} (L(n-1, k) + l(n, m))
\]
\[
\Psi(n, m) = \arg \max_{k=m,m-1} L(n-1, k),
\]
\[
\psi(n) = \Psi(n + 1, \psi(n - 1)).
\]

To guarantee the increasing order of the resultant stretching factors, the following constraints are imposed in a Viterbi search processing,
\[
\Psi(n, m) = \begin{cases} 
1 & \text{if } m = 1 \\
M & \text{if } m = n \\
m - 1 & \text{if } m > n \end{cases}
\]
\[
\psi(N) = M.
\]

The classification boundaries are determined by finding the nodes in the trellis where the direction of the path is changed, as shown in Fig. 2. Hence, the resulting set of the classification boundaries is given by
\[
\Theta_i = \{ \theta_i = r_{n-1} | \psi(n - 1) \neq \psi(n), 1 \leq n \leq N \}.
\]

Step-2. Determining the optimal stretching factors for each class: With a previously determined \(\Theta_i\), find the stretching factor for each class that maximizes the likelihood of the stretched cepstrum coefficients belonging to the corresponding class.
\[
\gamma^{(i+1)}_m = \arg \max_{\gamma} \sum_{x \in S_n} \log p(f(x, \gamma)|\Lambda), 1 \leq m \leq M.
\]

Step-3. Convergence Test: Compute the overall likelihood at the iteration \(i\) with \(\Theta^{(i)}\) and \(\Gamma^{(i+1)}\).
\[
\lambda^{(i)} = \mathcal{L}(X, \Lambda, \Theta^{(i)}, \Gamma^{(i+1)}).
\]

if \( (\lambda^{(i)} - \lambda^{(i-1)}) / \lambda^{(i-1)} \leq \epsilon \), stop with \(\Theta^{(i)}\) and \(\Gamma^{(i+1)}\) describing the final set of the classification boundaries and the final set of the stretching factors. Otherwise replace \(i\) by \(i + 1\) and proceed to Step-1.

Note that since it cannot be mathematically defined the relationship between \(\gamma\) and \(\log p(f(x, \gamma)|\Lambda)\) in (10), \(\gamma^{(i+1)}_m\) cannot be obtained analytically. Thus, the use of a full search algorithm which requires huge computation cannot be avoided. To alleviate this problem, we set limits on the number of possible stretching factors. The experimental results show that a stretching factor of less than 1.2 did not significantly affect the performance of the proposed ASR system, whereas excessive stretching factors significantly degraded the performance of the ASR system. Hence, we used a possible stretching factor set \(\mathcal{A} = \{1.25, 1.3, 1.35, 1.4, 1.45, 1.5, 1.55, 1.6, 1.65, 1.7, 1.75, 1.8, 1.85, 1.9\}.

2.3 Stretching the Cepstrum Coefficients

To produce longer cepstrum sequence than original one, the intermediate cepstrum vectors between the neighboring samples should be estimated. In [7], the intermediate samples were estimated in three different ways. Among them, approximated bandlimited interpolation revealed the best performance, where Lanczos filter is employed. In the present work, a similar method is employed, where the bilinear interpolation is carried on the incoming cepstrum coefficients.

3. Experimental Results

ASR experiments were carried out on 10-digit mobile phone numbers, pronounced in Korean. Speech samples were recorded from 16 speakers, 8 speakers for training, the remaining 8 speakers for testing. Each speaker pronounced 10 species of phone numbers under noise-free conditions. The speech corpus was composed of two sets: a speech set for training, a speech set for testing. Each speech set contained 1600 utterances, which corresponds to 19200 words.

Each utterance is lowpass filtered up to 7.5 kHz. The sampling rate is 16 kHz. The feature parameters consist of 13 MFCCs (Mel Frequency Cepstrum Coefficients), 13 delta MFCCs and 13 delta-delta MFCCs. Hence, the total number of acoustic feature parameters for defining HMMs is 39. A 25 msec length Hanning window was used to compute and extract the MFCCs at 10 msec intervals. A preemphasis factor 0.95 was applied and the number of mel-frequency filter banks were 24. The speech recognizer uses triphone-based HMMs with three states for each triphone. There are 14 triphones in the database. The distribution of acoustic features were modelled using mixtures of diagonal Gaussians.

To evaluate the effectiveness of the proposed method in ASR, it is necessary to build a reference model for fast speech. To this end, we further split the training speech set into two subsets. For each subset, we then individually trained two HMMs, that is, HMM-N and HMM-F, for use as reference models for a normal rate speech and for fast speech, respectively. Splitting was performed by comparing the estimated ROS with its median value.

The results are shown in Fig. 3. The proposed method always gives better performance than the ASR method where one HMM from training speech set is used. The recognition accuracy is increased as the number of classes is increased (except when the number of classes is 14). The highest performance of the proposed ASR system was achieved when the number of classes is 20. The overestimation-problem was encountered in the case of using more than 20 classes. This resulted in inconsistent performance between training and test utterances.

It is noteworthy that the proposed method has 17.8% fewer errors for test utterances in case when the number of classes is 20. Compared with the method where multiple specialized HMMs by ROS is used (top line), the av-
Average word recognition rate is slightly reduced. The absolute difference in word recognition rate between the two methods is 1.8% (91.7% vs. 89.9%, when the number of classes is 20). There can be several reasons for this difference. One is that the differences between normal rate speech and fast speech cannot be completely compensated for by time-stretching. Recent work [8] also noted that the characteristics of fast speech are more or less different from normal rate speech, especially in formant frequency contours. Therefore, it would be desirable that additional processing is carried on the time-stretched cepstrum coefficients, which will further increases the accuracy of the ASR systems.

4. Conclusion

We proposed a robust ASR system, especially for fast speech. There can be many ways to compensate for the differences between normal rate speech and fast speech in terms of ASR performance. Our approach focused on modifying the entire duration of the incoming utterances according to their rate. The major contribution of the present work includes how the incoming utterances are classified in accordance with ROS and how the stretching factors for each class are determined. From our experimental results, improvements in the recognition rate were achieved by the proposed method. This indicates that the proposed method can be successfully applied to existing ASR systems, especially for increasing the robustness of fast speech.

Other modification methods, including pitch modification can also be useful for compensating for adverse conditions in ASR. Future work will focus on this issue.

References