Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 48 (2006) 1100–1111
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

HMM-based noise-robust feature compensation

Akira Sasou [a,*], Futoshi Asano [a], Satoshi Nakamura [b], Kazuyo Tanaka [a,c]

[a] *National Institute of Advanced Industrial Science and Technology (AIST), Information Technology Research Institute, 1-1-1 Tsukuba, Ibaraki 305-8568, Japan*
[b] *Advanced Telecommunications Research Institute International (ATR), Spoken Language Translation Research Laboratories, Japan*
[c] *University of Tsukuba, Institute of Library and Information Science, Japan*

Received 27 April 2005; received in revised form 28 November 2005; accepted 13 March 2006

Abstract

In this paper, we describe a hidden Markov model (HMM)-based feature-compensation method. The proposed method compensates for noise-corrupted speech features in the mel-frequency cepstral coefficient (MFCC) domain using the output probability density functions (pdfs) of the HMM. In compensating the features, the output pdfs are adaptively weighted according to forward path probabilities. Because of this, the proposed method can minimize degradation of feature-compensation accuracy due to a temporarily changing noise environment. We evaluated the proposed method based on the AURORA2 database. All the experiments were conducted under clean conditions. The experimental results indicate that the proposed method, combined with cepstral mean subtraction, can achieve a word accuracy of 87.64%. We also show that the proposed method is useful in a transient pulse noise environment.

Keywords: Noise robust; Hidden Markov model; AURORA2

1. Introduction

One approach to noise robust speech recognition is to use a microphone array which can enhance a speech signal contaminated by directional noise sources. A spatial inverse approach forms a valley of sensitivity in a certain direction by means of adaptation/learning. However, for omni-directional noise sources, the spatial inverse approach tends to be less effective. In addition, because, in the adaptation/learning phase, sufficient data are required in order to estimate stable statistics such as a spatial correlation matrix, the spatial inverse approach suffers from suppression of transient noises such as pulse noises.

In the case of use of a single microphone, model adaptation is an effective approach to robust speech recognition which adapts acoustic models to noisy conditions (Gales and Woodland, 1996; Varga and Moore, 1990; Gales and Young, 1996). Model adaptation methods have an advantage in that these methods can adapt not only expectations, but also variances. When the noise environment can be modeled in advance, methods such as parallel model

* Corresponding author. Tel.: +81 29 861 5755; fax: +81 29 861 5939.
E-mail addresses: a-sasou@aist.go.jp (A. Sasou), f.asano@aist.go.jp (F. Asano), satoshi.nakamura@atr.jp (S. Nakamura), ktanaka@ulis.ac.jp (K. Tanaka).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2006.03.002

combination are useful for adapting acoustic models. However, when the methods are applied to an unknown noise environment, it is difficult to maintain sufficient recognition accuracy.

Another effective approach to realize noise-robust speech recognition is feature compensation of noise-corrupted speech during the front-end processing (Boll, 1979; Atal, 1974). Segura et al. (2001) demonstrated the effectiveness of the feature-compensation method based on a Gaussian mixture model (GMM), which estimates the expectation of distortion from a noise-corrupted speech feature based on the GMM in the log filter bank energy (log-FBE) domain, and generates a compensated speech feature by subtracting the estimated distortion from the corrupted speech feature. The GMM is prepared in advance by being trained from clean-speech features. In the feature-compensation process, the GMM is adapted to the noisy environment using several frames prior to an utterance, which are expected to be noise-only frames. Because the adapted GMM is well matched to noise-corrupted speech features, the method can adequately compensate for the corrupted speech features in a stationary noise environment. However, in the case of a non-stationary noise environment, because the noise characteristics change temporally, mismatches between the adapted GMM and a current noise-corrupted speech feature occur. Because of this mismatch, the posterior probabilities of the Gaussian distributions, which should not contribute to compensation for features, are over-estimated. This causes degradation of the feature-compensation accuracy. Another drawback is that the feature-compensation formula of the GMM-based method cannot be applied to noise environments that have no stationary noise component, such as transient pulse noises.

To achieve robust speech recognition in a non-stationary noise environment, Yao et al. (2000) proposed an effective method that separates the noise effects into stationary and residual, and estimates the residual noise parameters at each time with a sequential EM algorithm. However, the computation cost is very high because the method needs a number of iterations at each time in order to converge the residual noise parameters. Recently, in order to reduce the computation cost, a particle filter-based sequential estimation method has been adopted in some methods, e.g. Yao and Nakamura (2001); Fujimoto and Nakamura (2005).

In this paper, we propose a hidden Markov model (HMM)-based feature-compensation method that can deal with, in addition to the stationary noises, non-stationary noises that temporarily vary their characteristics like transient pulses. The proposed feature compensation is achieved based on the HMM that the decoder has in advance as an acoustic model, and is thus executed in a cepstral domain. Moreover, because the proposed feature compensation is applicable to the residual noise after the microphone array processing, the integration system of the proposed feature compensation and the microphone array is expected to make it possible to cover a wide variety of noise. A distortion of the noisy speech feature in the cepstral domain, which is different from the clean speech feature, can be divided into a stationary distortion component and a non-stationary distortion component. The temporal trajectory of the non-stationary distortion component is assumed to be zero almost everywhere but temporarily changes. The stationary distortion component is absorbed by adding the estimated stationary distortion component to the expectation value of each Gaussian distribution in the output probability density functions (pdfs) of the clean speech's HMMs. The proposed method eliminates the degradation of feature-compensation accuracy caused by the non-stationary distortion component by evaluating the posterior probability of each Gaussian distribution that is adaptively weighted based on the forward path probability. Even if the current noise characteristics have changed temporarily, the proposed method can enhance the posterior probabilities of the Gaussian distributions, which should contribute to the feature compensation.

The idea of separating the noise effect into stationary and residual parts was also adopted by Yao et al. (2000). However, our method does not need to estimate the residual noise parameters using such a complex processing as the sequential EM algorithm or the particle filter-based method, and simply applies weights to the pdfs in order to compensate for the residual noise effects in evaluation of posterior probabilities of Gaussian distributions. The weights are evaluated based on the forward path probabilities of the Viterbi algorithm. Our method is thus very simple and can easily be implemented with an existing speech recognizer, which is one of our method's advantages. In order to overcome another drawback of the GMM-based method, we also propose a feature-compensation formula that can be applied to noise environments that have no stationary noise component, such as transient pulse noises.