Research and Development of Robust Speech Recognition

Kiyoaki Aikawa

NTT Human Interface Laboratories, 1-1 Hikarino-oka, Yokosuka-shi, Kanagawa 239 Japan

Abstract

This paper describes recent research and development activities on robust ASR (automatic speech recognition) in NTT Human Interface Laboratories. ASR system design has been changing from the experimental to the commercial level. A relevant issue in achieving practical ASR is robustness against environmental noise and speaker/circuit differences. Adaptation techniques have been widely investigated for improving the robustness. Real time speaker/noise adaptation is getting more and more popular. Confidence measures are also effective in enhancing speech recognition performance by rejecting out-of-vocabulary words, lip and breath noises. Another relevant issue is how to quickly reach the goal of human-machine dialogue via the telephone without visual information.

INTRODUCTION

The development of ASR techniques has been accelerating in recent years. The progress of computer technology greatly contributes to the research and development of ASR. It is also an advantageous condition that even a standard PC is equipped with AD and DA converters. This paper reports recent ASR research and development activities in NTT Human Interface Laboratories focusing on robustness enhancement.

ACOUSTIC MODEL

Mixture Gaussian models have been used more often in recent HMMs (hidden Markov models). The probability distribution represented by a large number of Gaussians actually provides good performance in speaker-independent speech recognition. The question is does the mixture Gaussian offer the best way of representing a complex probability distribution of acoustical feature or not. Discrete mixture HMM is a new idea for representing the distribution of acoustic feature parameters in HMM states [7]. The probability density function is represented by a histogram for each feature parameter such as cepstral coefficient. The overall distribution is represented as a mixture of the individual density functions. The proposed model reduced the computation time under the condition of fixed recognition accuracy. Each feature parameter value is quantized into a fixed number of sections which covers the entire range of the feature parameter distribution.

If out-of-vocabulary and lip/breath noises may be included in the input speech, rejecting out-of-vocabulary words provides robust speech recognition by accepting only a set of valid keywords. A new confidence measure incorporates likelihood differences among phonetic HMMs [3]. The proposed measure can be embedded in the search process of ASR.

ADAPTATION

HMM composition is a noise adaptation technique that generates the acoustic model representing noise-added voice spectra [4]. Jacobian noise adaptation has been developed for adapting from one environmental noise to a different environmental noise based on the HMM composition technique [8]. The method uses a Jacobian matrix whose elements are the first order Taylor-expansion coefficients representing the change of cepstral coefficient over the change of noise cepstrum. Jacobian adaptation needs only a very short segment of noise data, less than 1 sec. The training time was reduced to as low as one 34th of that by the conventional HMM composition technique [4]. Fast adaptation is strongly desired especially for achieving network services.
such as information retrieval via telephone, because the system has to collect a noise sample in a short period before the user begins to talk to the system.

It is usually difficult to collect a large amount of speech data spoken by a particular speaker. Thus most existing ASR systems use speaker-independent acoustic models. To provide better speech recognition performance speaker adaptation techniques have been investigated. A new speaker adaptation method was introduced based on N-best word sequence hypotheses [5]. Intermediate HMM parameters are trained for each of the N-best hypotheses. The final adapted HMM parameters are obtained as the confidence-measure-weighted sum of these intermediate parameters.

Unsupervised speaker adaptation of HMMs is convenient for automatic improvement of ASR performance. A newly proposed unsupervised adaptation method uses the likelihood distribution of top hypotheses [1]. The advantage of this technique is to automatically improve the recognition performance in batch dictation of large amount of speech data. The threshold deciding the availability of speech data for training was determined based on the signal detection theory. The method is based on a fact that the likelihood difference between word sequence hypotheses is generally large if the top candidate is correct.

**TOPIC EXTRACTION**

Speech understanding is being focused on in the new research field of broadcast news recognition. An application of speech understanding is news database handling such as searching or indexing news stories. A new relationship function has been derived for representing the closeness between a keyword set and a news topic [6]. The relationship function is formulated based on the chi-square value. The new function improved topic extraction accuracy for broadcast news dictated by an ASR system.

**DIALOGUE CONTROL**

It is necessary to prevent a human-machine dialogue from falling into an infinite cycle of Q/A or diverging away from the desired target because of speech recognition error. A new dialogue optimization method was proposed based on the minimization of average number of Q/As issued to complete a dialogue [2]. Several subordinating properties are assumed to be attached to each items. For example, each of three entry items has two subordinate information, category 1 = { a, p, e } and category 2 = { bat, rat, dog}. To decide between item 1 or 2 at a Q/A in a dialogue flow, the use of category 1 { a/p } will be better to use in minimizing recognition error. However, to decide between item 2 and 3, category 2 { rat/dog } will be better. A measure for estimating the number of Q/As was derived based on the phoneme confusion matrix and phonetic transcription of words.

**SUMMARY**

ASR techniques have progressed, however, good performance is achieved only for clear utterance under high quality environmental conditions. ASR techniques are still weak against non-stationary noise, unknown word inserted between words, repeated utterance of the same word or utterance correction. Future work should focus on these problems. Intelligent error recovery is another issue for building a more comfortable and more practical ASR application system.

**References**