STEPS TOWARD FLEXIBLE SPEECH RECOGNITION

– RECENT PROGRESS AT TOKYO INSTITUTE OF TECHNOLOGY –

Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

ABSTRACT - This paper describes recent progress at Tokyo Institute of Technology and the author's perspectives for making speech recognition systems more flexible at both the acoustic and linguistic processing levels. Specifically, it describes a broadcast news transcription system, a multimodal dialogue system for information retrieval, neural-network-based HMM adaptation for noisy speech, online incremental speaker adaptation combined with automatic speaker-change detection, message-driven speech recognition and understanding, speech summarization, a Japanese national project on spontaneous speech corpus and processing technology, and speech recognition in the ubiquitous/wearable computing environment. For processing spontaneous speech, indispensable will be a paradigm shift from speech recognition to understanding where underlying messages of the speaker are extracted, instead of transcribing all the spoken words. Building a large corpus of spontaneous speech to construct reliable acoustic and linguistic models is also crucial. Due principally to the technology of making computers smaller, more powerful and cheaper, the ubiquitous and wearable computing era is expected to come into being in the initial years of the 21st century. In such an environment, speech recognition will be widely used as one of the principal methods of human-computer interaction.

Keywords - Flexible speech recognition, model adaptation, speech understanding, speech summarization, spontaneous speech, ubiquitous/wearable computing environment.

INTRODUCTION

The field of automatic speech recognition has witnessed a number of significant advances in the past 5-10 years, spurred on by advances in signal processing, algorithms, computational architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, a data-driven approach which makes use of a rich set of speech utterances from a large population of speakers, the use of stochastic acoustic and language modeling, and the use of dynamic programming-based search methods.

Figure 1 shows a mechanism of state-of-the-art speech recognizers [1]. Common features of these systems are the use of cepstral parameters and their regression coefficients as speech features, triphone HMMs as acoustic models, vocabularies of several thousand or several tens of thousand entries, and stochastic language models such as bigrams and trigrams. Such methods have been applied not only to English but also to French, German, Italian and Japanese. Although there are several language-specific characteristics, similar recognition results have been obtained.

Ultimately, speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition. Figure 2 shows the main causes of acoustic variation in speech [2]. It is crucial to establish methods that are robust against voice variation due to individuality, the physical and psychological condition of the speaker, telephone sets, microphones, network characteristics, additive background noise, speaking styles, and other aspects. Also important is for the systems to impose few restrictions on tasks and vocabulary. Developing automatic adaptation techniques is essential to resolve these problems.

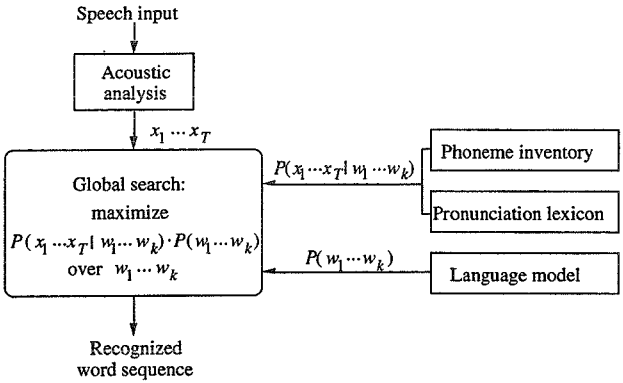

Fig. 1 - Mechanism of state-of-the-art speech recognizers.

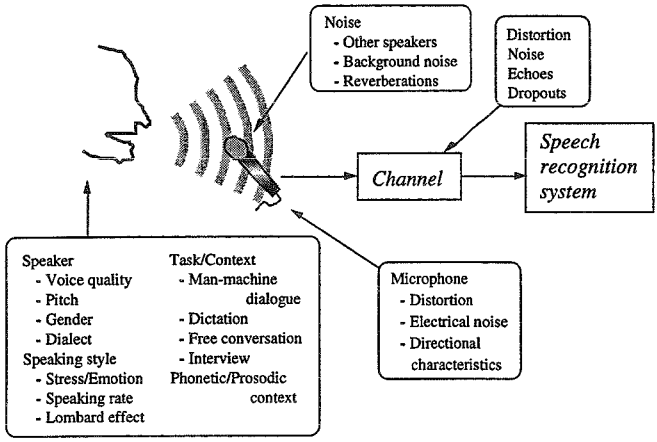

Fig. 2 - Main causes of acoustic variation in speech.

Another important issue for speech recognition is how to create language models (rules) for spontaneous speech. When recognizing spontaneous speech in dialogues, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include extraneous words, out-of-vocabulary words, ungrammatical sentences, disfluency, partial words, repairs, hesitations, and repetitions. It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. A paradigm shift from the present transcription-based approach to a detection-based approach will be important to resolve such problems [3]. How to extract contextual information, predict users' responses, and focus on key words are very important issues.

Stochastic language modeling, such as bigrams and trigrams, has been a very powerful tool, so it would be very effective to extend its utility by incorporating semantic knowledge. It would also be useful to integrate unification grammars and context-free grammars for efficient word prediction. Style shifting is also an important problem in spontaneous speech recognition. In typical laboratory experiments, speakers are reading lists of words rather than trying to accomplish a real task. Users actually trying to do the latter, however, use a different linguistic style. Adaptation of linguistic models according to tasks, topics and speaking styles is a very important issue since collecting a large linguistic database for every new task is difficult and costly.

BROADCAST NEWS TRANSCRIPTION

We have been developing a Japanese broadcast-news speech transcription system [4] as part of a joint research project with NHK broadcasting. The goal is the closed-captioning of TV programs. The language models were constructed using broadcast-news manuscripts taken from NHK TV news broadcasts made between July 1992 and May 1996. The broadcasts comprised roughly 500k sentences and 22M words (morphemes). To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words using a morphological analyzer because Japanese sentences are written without spaces between words. Since many Japanese words have multiple readings and correct readings can only be decided according to context, we constructed a language model in which a word with multiple readings is split into different language model entries according to those readings. We also introduced filled-pause modeling into the language model. A word-frequency list was derived for the news manuscripts, and the 20k most frequently used words were selected as vocabulary words. This 20k vocabulary covered approximately 98% of the words in the manuscripts. We calculated bigrams and trigrams and estimated unseen n-grams using Katz's back-off smoothing method.

The feature vector consisted of 16 cepstral coefficients, normalized logarithmic power, and their delta features (regression coefficients). The total number of parameters in each vector was 34. Cepstral coefficients were normalized by the cepstral mean subtraction (CMS) method. The acoustic models were gender-dependent shared-state triphone HMMs and were designed using tree-based clustering. They were trained using phonetically-balanced sentences and dialogues read by 53 male speakers and 56 female speakers. The contents were completely different from the broadcast-news task. The total number of training utterances was 13,270 for the males and 13,367 for the females, and the total length of the training data was approximately 20 hours for each gender. The total number of HMM states was approximately 2,000 for each gender, and the number of Gaussian mixture components per state was four.

Clean speech data consisting of 50 male and 50 female utterances with no background noise were extracted from TV news broadcast in July 1996 and used as evaluation utterances. The male and female sets included utterances by five or six speakers, respectively. All utterances were manually segmented into sentences. The mean word error rates for male and female utterances were 14.2% and 12.9%, respectively.

DESIGNING A MULTIMODAL DIALOGUE SYSTEM FOR INFORMATION RETRIEVAL

We have investigated a paradigm for designing multimodal dialogue systems [5]. An example task of the system was to retrieve particular information about different shops in the Tokyo Metropolitan area, such as their names, addresses and phone numbers. The system accepted speech and screen touching as input, and presented retrieved information on a screen display or by synthesized speech as shown in Fig. 3. The speech recognition part was modeled by the FSN (finite state network) consisting of keywords and fillers, both of which were implemented by the DAWG (directed acyclic word-graph) structure. The number of keywords was 306, consisting of district names and business names. The fillers accepted roughly 100,000 non-keywords/phrases occuring in spontaneous speech. A variety of dialogue strategies were designed and evaluated based on an objective cost function having a set of actions and states as parameters. Expected dialogue cost was calculated for each strategy, and the best strategy was selected according to the keyword recognition accuracy.

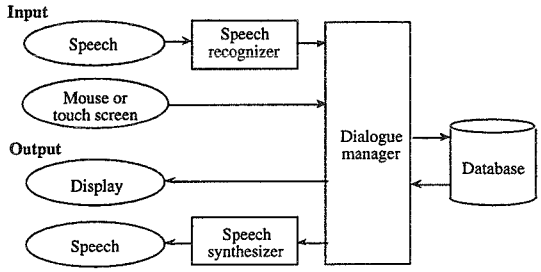

Fig. 3 - Multimodal dialogue system structure for information retrieval.

NEURAL-NETWORK-BASED HMM ADAPTATION FOR NOISY SPEECH

Increasing the robustness of speech HMMs (hidden Markov models) in terms of additive noise is one of the most important issues in state-of-the-art speech recognition. HMMs are usually represented by Gaussian mixtures in the multi-dimensional space of cepstral coefficients, in other words, by parameters in the logarithmic spectral domain. However, noise is added to speech in the waveform or in the linear spectral domain, so the incorporation of additive noise into HMMs is not straightforward. Parallel model combination (PMC, also called HMM composition) [6][7] is one of the most useful methods used to handle additive noise. PMC can be used to derive noisy speech HMMs by combining clean speech HMMs, a noise HMM and a signal-to-noise ratio (S/N). However, this method requires nonlinear conversions of the distribution parameters between cepstral and linear spectral domains, and also some approximations in the computational process.

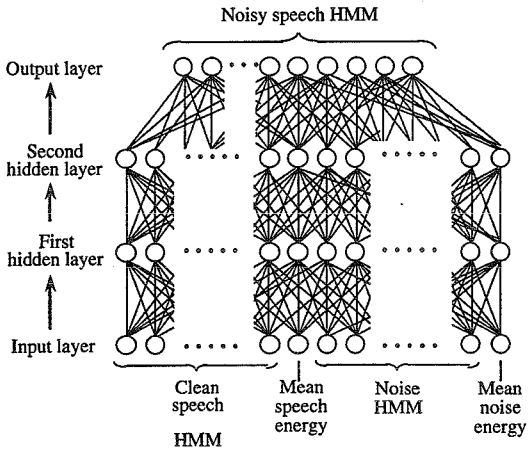

Fig. 4 – Structure of neural networks used for HMM noise adaptation.

We have proposed a method using mapping functions of neural networks to represent the total effect of additive noise on HMMs including nonlinear computations [8]. In this method, the neural networks are trained to map clean speech HMMs to noise-added speech HMMs, using noise HMM and S/N ratio as additional input signals, as shown in Fig. 4. The noise-added speech HMMs used as the target/ideal output signals for the neural networks training are made using noise-added speech produced by summing various clean speech and noises with various S/N ratios. Once the network is trained under various conditions of speech, noise and S/N, the network is expected to produce noise-added speech HMMs under a new condition of speech, noise and S/N within the generalization capability of the neural networks. In the present framework, only the mean vectors of Gaussian mixtures are adapted. Covariance values are preserved unchanged for simplicity.

The network is trained to minimize the mean squared error between the output HMMs and the target noise-adapted HMMs. Noisy broadcast-news speech was recognized in speaker-dependent and speaker-independent network training conditions, and the trained networks were confirmed to be effective in the recognition of new speakers and under new noise and S/N conditions.

ONLINE INCREMENTAL SPEAKER ADAPTATION

Extraction and normalization of (adaptation to) voice individuality is one of the most important issues in robust speech recognition. A small percentage of people occasionally cause systems to produce exceptionally low recognition rates. This is an example of the "sheep and goats" phenomenon. Speaker adaptation (normalization) methods can usually be classified into supervised and unsupervised methods. Unsupervised, online, instantaneous/incremental adaptation is ideal, since the system works as if it were a speaker-independent system, and it performs increasingly better as it is used. However, since we have to adapt many phonemes using a limited size of utterances including only a limited number of phonemes, it is crucial to use reasonable modeling of speaker-to-speaker variability or constraints.

In many applications of speech recognition, speakers change frequently, new speakers appear, and each of them utters a series of several sentences. In such a situation, an unsupervised and online adaptation method, which uses the unknown utterance itself for adaptation, is expected to be effective. The adaptation should also work incrementally within a segment in which one speaker utters several sentences. To create such a system, we must ensure that speaker change is detected automatically and correctly.

We have proposed an online, unsupervised, instantaneous and incremental speaker adaptation method combined with automatic detection of speaker changes for broadcast news transcription [9]. The MLLR [10] -MAP [11] and VFS (vector-field smoothing) [12] methods were instantaneously and incrementally carried out for each utterance. The speaker change is detected by comparing likelihoods using speaker-independent and speaker-adaptive GMMs (Gaussian mixture models). A single-state speaker-independent GMM with 64 Gaussian distributions is constructed in the training stage using the same utterances that were used to construct the speaker-independent phone HMM.

The adaptation process is shown in Fig. 5. For the first input utterance, the speaker-independent model is used for both recognition and adaptation, and the first speaker-adapted HMM and GMM are created. For the second input utterance, the likelihood value of the utterance given the speaker-independent GMM and that given the speaker-adapted GMM are calculated and compared. If the former value is larger, the utterance is considered to be the beginning of a new speaker, and another speaker-adapted HMM as well as GMM are created. Otherwise, the existing speaker-adapted HMM and GMM are incrementally adapted. For the succeeding input utterances, speaker changes are detected in the same way by comparing the acoustic likelihood values of each utterance obtained from the speaker-independent GMM and speaker-adapted GMMs. If the speaker-independent GMM yields a larger likelihood than any of the speaker-adapted GMMs, a speaker change is detected and new speaker-adapted HMM and GMM are constructed. Since GMM has only one state, it is adapted using a transformation matrix obtained by a single-cluster MLLR adaptation procedure applied to the phone HMM although HMM is actually adapted using seven phonemic clusters.

The same utterances used in the experiments described in "Broadcast News Transcription" were used in the evaluation. Experimental results show that the adaptation reduced the word error rate by 10.0% relative to the speaker-independent models.

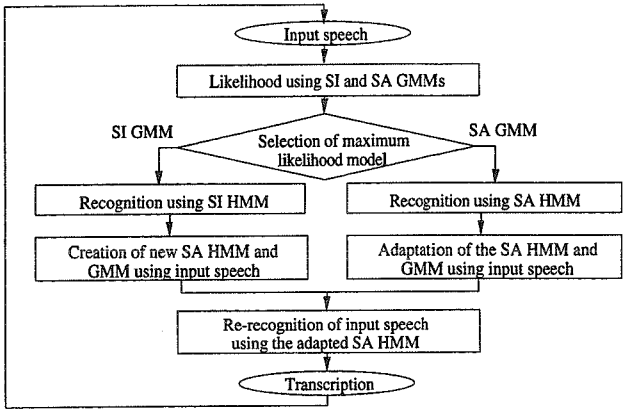

Fig. 5 – Online incremental speaker adaptation process including automatic speaker-change detection (SI: speaker independent; SA: speaker adapted).

MESSAGE-DRIVEN SPEECH RECOGNITION AND UNDERSTANDING

State-of-the-art automatic speech recognition systems employ the criterion of maximizing $P(W|X)$, where W is a word sequence, and X is an acoustic observation sequence. This criterion is reasonable for dictating read speech.

However, the ultimate goal of automatic speech recognition is to extract the underlying messages of the speaker from the speech signals. Hence we need to model the process of speech generation and recognition as shown in Fig. 6 [13], where M is the message (content) that a speaker intended to convey. There is also a possibility to give feedback from the "understanding module" to the speech recognition module such that decoding hypotheses can be properly adjusted and, hopefully, converge to the most correct word sequence as well as the most correct understanding of the utterance.

According to this model, the speech recognition process is represented as the maximization of the following a posteriori probability [4][13]:

$$\max_M P(M \mid X) = \max_M \sum_W P(M \mid W)P(W \mid X). \tag{1}$$

Using Bayes' rule, Eq. (1) can be expressed as

$$\max_M P(M \mid X) = \max_M \sum_W \frac{P(X \mid W)P(W \mid M)P(M)}{P(X)}. \tag{2}$$

For simplicity, we can approximate the equation as

$$\max_M P(M \mid X) = \max_{M,W} \frac{P(X \mid W)P(W \mid M)P(M)}{P(X)}. \tag{3}$$