21 年 01 月 19 日

| Electronics and Information Engg. Department | ID | 069301 |
|---|---|---|
| Name | Huda Mohammad Nurul | |

| Advisor | Tsuneo Nitta<br>Kouichi Katsurada |
|---|---|

| Title | A study on articulatory feature extraction for robust speech recognition |
|---|---|

(800 words)

Automatic speech recognition (ASR) research efforts have been intensified over the last decade. Because of the development of both efficient speech recognition algorithms and powerful hardware, the quality of ASR systems has increased drastically. ASR systems, which are currently available, have been developed for a wide variety of applications. Nevertheless, speech recognizers are still far from being ubiquitous. The reason why speech recognition has not found more widespread use and why its commercial potential has not been fully exploited is the lack of robustness, against noise, speakers' variability, and coarticulation, of current ASR technology in practical environments.

Various hidden Markov model (HMM)-based ASR systems have been developed. Most of these ASR systems make use of a preprocessed form, such as mel-frequency cepstral coefficient (MFCC), of the speech signal, which encodes the time-frequency distribution of signal energy. The goal of this thesis is to investigate the benefits of integrating articulatory information into state-of the art speech recognizer as a genuine alternative to standard acoustic representations. This articulatory information describes properties of speech production rather than the properties of acoustic signal. Articulatory information is represented in term of articulatory classes or "features", which are called distinctive phonetic features (DPFs).

A phoneme can easily be identified by using its unique DPF set, which comprises the manner of articulation (vocalic, consonantal, continuant, etc.) and place of articulation (tongue position: high, low, front, back, etc). The use of DPFs in ASR had been investigated previously, and has been actively discussed in recent years. However, DPF isn't widely used as features for ASR instead of MFCC because DPF itself cannot provide enough performance. This thesis presents a method to extract DPFs using two stages of multilayer neural networks (MLNs) and to apply the DPFs to noise-robust ASR. Since DPFs are designed after full consideration of speech production, a DPF-space well represents the distances among phonemes corresponding to their articulation differences. This fact suggests that DPFs are efficient feature parameters for ASR. In the DPF extractor construction, the first MLN takes a preceding, current, and following context acoustic vectors as input and outputs three corresponding context-dependent DPF vectors. On the other hand, the second MLN outputs three DPF context-dependent DPF vectors, which exhibit reduced context effects, by taking the output of the first MLN and corresponding delta and delta-delta parameters as input.

Though this two-stage MLNs based DPF extractor provides higher phoneme recognition performance under clean and noisy environments over the extractor designed by a single MLN, it is needed to achieve more categorical DPF movement by enhancing DPF peaks (convex patterns) and by inhibiting DPF dips (concave patterns) for obtaining more correctness. In this thesis, we introduced a robust algorithm, which is called Inhibition/Enhancement network, for obtaining

-Continue-

better DPF patterns. This algorithm develops an inhibitory or an enhancement function based on the DPF movement and provides modified DPF patterns by multiplying the original DPF patterns, produced by the DPF extractor, with the inhibition/enhancement function.

On the other hand, the modified DPFs extracted from Inhibition/Enhancement network correlate between components. Since each of these three context-dependent DPF vectors is not orthogonalized to each other, they should be decorrelated with respect to the current context vector using the Gram-Schmidt orthogonalization procedure before connecting with an HMM-based classifier.

A DPF-based phoneme recognizer performance, both phoneme correctness and accuracy, can be enhanced by incorporating syllable-based subword language models. These syllables set grammatical rules to retrain the violation of phonotactic constraints at the acoustic phonetic level. A language model is adopted in a recognition system for post-processing phoneme estimates and making correction with the grammatical constraints. In this thesis, more accurate phonemes string for an input speech are generated by incorporating acoustic and language models with DPF extractor based on three stage MLNs and Inhibition/Enhancement network.

Acoustic models (AMs) of the HMM-based classifier include various types of hidden variables such as gender type, speaking rate, and acoustic environment. They also affect the DPF-based ASR system. Therefore, if there exists a canonicalization process that reduces the influence of the hidden variables from AMs, a robust ASR system can be realized. In this thesis, finally, the configuration of the canonicalization process targeting gender type and noise intensity as hidden variables is addressed. The canonicalization process is realized by introducing a DPF space between an acoustic feature space and AMs of the HMM classifier. The proposed canonicalization process is composed of multiple DPF extractors corresponding to the hidden variables canonicalization, and a DPF selector which selects an optimum DPF vector from multiple DPF vectors an input of the HMM-based classifier. On the other hand, noise factor can be eliminated by applying canonicalization based on the DPF extractors and two-stage Wiener filtering. Each of the DPF extractor of this process is implemented by a single MLN. The canonicalization process of feature parameters through the DPF space reduces improper influence of the hidden variables.