学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　Federico Mendoza Ang

学　位　論　文　題　名

A Study of Noise-Robust Speech Recognition and Time-Varying Speech Features
（時変音声特徴抽出とロバスト音声認識に関する研究）

Noise or variabilities in speech signals take different predominant forms across different languages. Two case scenarios are presented, one where the variability can be solved using data-driven methods and one where a knowledge-based hypothesis is necessary. In a data-driven solution, the compensation for mismatches is supplied using more data that is representative of the missing information. Such is the case for languages that make use of code-switching, or words from another language as a form of variability.

For noise problems requiring a knowledge-based solution, the first line of defense is with the speech features, as it represents the only basis for the entire recognition system. Because of strict requirements in real-time processing for practical applications, the development of speech features for automatic speech recognition systems were driven by short-time analysis methods. A clear limitation of short-time analysis is it fails to capture fast-changing phenomena within a frame of speech as it computes a single output representation. The limitation is further supported by the augmentation of derivative-based features to capture transitions and acquire gains in performance.

As a solution, this work provides a feature extraction framework in which time-varying features are provided that is equal to the number of samples rather than a single representation per frame. Experiments were conducted on a highly acoustic model dependent speech recognition task to reveal issues from analysis of results. It is concluded that in its basic formulation, gains can be acquired by limiting the time-varying extraction only to frames that require accurate modeling, such as signal onsets. This finding results to hybrid systems combining time-varying and time-invariant features that can improve the baseline recognition rate for up to an average of 2% including noisy environments.

This thesis is divided into seven chapters beginning with a discussion of the importance of speech recognition systems, a short summary of the conducted research in the thesis, the research background and its objectives. Chapter 2 then provides a different case scenario that highlights the common approach to dealing with errors due to variability in speech recognition. In particular, this chapter showcases the development of a Large Vocabulary Continuous Speech Recognition (LVCSR) system where typically, data-driven methods are used. The discussion in this chapter will serve as a contrast to a knowledge-driven method that will become the main focus of the thesis. Through the experiments made in this chapter, the importance of knowledge-driven methods alongside data-driven methods becomes apparent. Chapter 3 gives an overview of isolated-word speech recognition systems. It hinges

on the common notion that speech recognition systems are highly application dependent. The aim of this chapter is to present the baseline architecture of the isolated-word speech recognition system that serves as the backbone for the knowledge-based focus of this thesis. The highlight of this chapter is in the feature extraction and acoustic modeling procedures used for the system. Aside from discussing the details of the working architecture, this chapter also summarizes the limitation and the currently known performance of the said baseline system, including the noise compensation techniques employed. This allows for a shift of the discussion to the proposed solution. Chapter 4 details the reasoning behind the problem of the baseline system, specifically with confusion between words that only differ by one or two phonemes, attributing the errors to the limitations of short-time based analysis of speech signals. Addressing the said limitations leads to the use of the proposed time-varying speech features, specifically based on the Linear Predictive Coding (LPC) cepstrum. This leads to the discussion of the feature extraction method proposed, the time-varying LP cepstral coefficients (TV-LPCC). The various implementations based on different criteria were also discussed, as well as possible issues, used solutions, and justifications made. These issues pertain to the stability of the system that is solved using replication and pole reflection, and numerical underflow for acoustic modeling that is solved using feature reduction. These reduction schemes serve as the main variations that are used for the experiments that follow. Chapter 5 deals with the practical aspects of setting up the experimental environment for conducting a reliable evaluation of the different models. This includes system configuration parameters, details of the database and its division, and creation of noisy models for robustness tests. Confidence measures and evaluation metrics are explained, as well as bias-variance considerations through the use of a held-out cross-validation scheme. Chapter 6 presents the results of the experiments and model selection procedures made. Post-experiment analysis and additional work are also discussed in this chapter. The preliminary results show some potential for the proposed scheme, with suboptimal performance for the basic formulation. By analyzing the performance of the proposed methods to the problems being addressed, joint models were realized by combining the performance of the baseline and proposed schemes. Immediate gains of 1-2% were acquired through these joint models. Finally, Chapter 7 concludes the thesis with a summary of results, contributions, and recommendations for future work.

The results gathered in this thesis is important for the advancement of the acoustic modeling aspect of speech recognition. This is especially important for cases where the system can only rely on the observed signal itself, such as for isolated-word speech recognition that are often found in application-specific consumer products. The solutions presented using the proposed scheme only adds quadratic complexity as a function of the LPC order using the first power as the basis functions. While the basic methods involving empirical parameter settings were shown to be effective for speech recognition, additional studies and experiments can be made for further improvement of the proposed method.