SIGNIFICANCE OF SPARSE REPRESENTATION FOR SPEECH RECOGNITION AND SPEECH SYNTHESIS

A THESIS

submitted by

PULKIT SHARMA

for the award of the degree

of

DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

March 19, 2018

A wise knows the truth that, "I do nothing at all"!

Bhagavat Gita

To My Grandfather

Pandit Udho Ram Sharma ji

And To My Parents

Neelam Sharma and Pradeep Kumar Sharma

Declaration

I hereby declare that the entire work embodied in this thesis is the result of investigations carried out by me in the **School of Computing and Electrical Engineering, Indian Institute of Technology Mandi**, under the supervision of **Dr. Anil Kumar Sao**, and that it has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgments have been made wherever the work described is based on finding of other investigators.

Mandi, 175005
Date:

Pulkit Sharma

THESIS CERTIFICATE

This is to certify that the thesis titled **SIGNIFICANCE OF SPARSE REPRESENTATION FOR SPEECH RECOGNITION AND SPEECH SYNTHESIS**, submitted by **Pulkit Sharma**, to the Indian Institute of Technology, Mandi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Mandi, 175005

Date:

Dr. Anil Kumar Sao

(Ph.D Supervisor)

Acknowledgments

First and foremost I would like to thank my advisor, Dr. A. K. Sao, for his inspiration, guidance and incredible support. Thanks to my doctoral committee members - Prof. B. D. Chaudhary, Dr. B. S. Rajpurohit, Dr. S. K. Sharma and Dr. Manoj Thakur. They have inspired me greatly though their own research, and their willingness to answer my questions has helped me greatly over the years.

I thank Prof. H. A. Murthy of IIT Madras for providing me an opportunity to spend three months in her lab during the first year of my Ph D. Our interactions gave me an opportunity to deepen my understanding of the fundamentals of speech processing. I sincerely thank her for providing me an irreplaceable advice both on my research and my career.

I am indebted to Dr. P. Rajan for carefully reviewing various drafts of our manuscripts and many fruitful discussions with Dr. Dileep A. D. are also gratefully acknowledged. In my daily work I have been blessed with a friendly and cheerful group of fellow students at the MAS lab: Srimanta, Pravindra, Shaifu, Vibha, Anshul, Ajay and many more. They provided a friendly and cooperative atmosphere at work and also provided useful feedback and insightful comments on my work. I am forever thankful to my friend and colleague Vinayak Abrol for his frequent help and support. Additionally, I would like to thank Kuldeep and Deshraj for providing assistance with administrative tasks and network related issues.

I thankfully acknowledge the fellowship that supported my research: Department of Electronics and Information Technology, Ministry of Communications and Information Technology, Government of India.

I express my heartfelt gratitude to my lovely grandmother, Smt Kesari Devi, beloved parents, Smt. Neelam and Shri Pradeep, my sister Pallvi, my brother Mukul and my cousin Rajat for their unconditional love and support throughout my life. My mother has been an

inspiration throughout my life, she has always supported my dreams and aspirations. I'd like to thank her for all she is, and all she has done for me.

Nivedita is certainly one of the main contributors of all the best in my life, and I am intellectually indebted to her ideas and our conversations. I thank God for the extraordinary chance of having her as wife, friend and companion.

Pulkit Sharma

Abstract

In this thesis, sparse representation (SR) based signal processing is employed to derive features for speech recognition and footprint reduction of unit selection based speech synthesis (USS) systems. The objective in speech recognition is to convert a speech utterance into text, however, the objective in speech synthesis is to generate speech corresponding to given text. In this work, the SR in speech recognition is employed to effectively discriminate among different speech units, while that in USS systems is used to compress the speech corpus.

In SR based signal processing, a speech signal is decomposed into a dictionary and the corresponding representation, having a few significant coefficients. The use SR for a particular application is greatly influenced by the choice of the dictionary. The data belonging to two confusing classes may lie in overlapping subspaces, and thus a single dictionary may not effectively discriminate among them. Hence, in this work, class specific principal component analysis (PCA) based multiple dictionaries are used for tasks in speech recognition. Here, speech frames belonging to each speech class/unit are clustered into different clusters, and a sub-dictionary is learned for each cluster. Our experiments reveal that coefficients corresponding to intermediate principal components (PCs) result in more discrimination among confusing speech units. Thus, a transformation function known as weighted decomposition is employed to emphasize the discriminative information present in the middle PC of the PCA-based dictionary. The performance of proposed features is evaluated using continuous density hidden Markov model based classifiers for various speech units classification tasks.

The use of class specific dictionaries in the proposed SR based feature results in an increased computational complexity. In order to address this issue, we have proposed to use a deep sparse representation (DSR) based unified model to learn a single multi level dictionary for all the speech classes. The proposed DSR model alternate between a sparse and a

dense layer, and it has been observed that representations obtained at different sparse layers have complimentary information. This means that a set of speech classes confusing at one layer is discriminative at another layer, and vice versa. Thus, we propose to concatenate representations obtained at different sparse layers to derive the final feature representation for speech recognition. GMM-HMM and DNN-HMM systems are used to evaluate the performance of the proposed feature for various speech recognition tasks. Experimental studies reveal that the deep dictionary derived using the proposed DSR model outperform both single overcomplete dictionary and multiple sub-dictionaries. The issue of speech recognition in noisy environment is addressed by enhancing the noisy speech, before deriving features for speech recognition. In particular, we have proposed a novel SR based method for speech enhancement, which is based on the observation that, given an appropriate dictionary, it is easy to estimate SR for speech signal, as opposed to noise.

The objective of speech synthesis can be seen as contrary to the speech recognition, and the USS system results in the best quality of synthesized speech, as compared to the contemporary approaches. In USS systems, speech units from a pre-recored speech database are selected and concatenated to synthesize speech. Thus, the size of speech database in USS systems limits its use in low resource devices. In this thesis, SR based signal processing is explored to compress the speech database to be stored in USS systems. The proposed method reduce the size of speech corpus by storing significant coefficients of the sparse vector. It is also observed that the behavior of SR varies for different speech sounds (e.g., voiced, unvoiced etc.). Hence, for efficient compression, different number of significant coefficients of the SR are stored for different speech sounds. USS systems build using two Indian languages (Hindi and Rajasthani) are used to evaluate the performance of the proposed compression methods. It has been shown that multiple dictionaries learned for individual speech units result in better compression in USS systems.

In addition, discriminating ability of SR is used in a kernel sparse representation based classifier (KSRC) for speech emotion recognition, where a given speech sample is classified into various categories of emotions. Further, a group sparsity constraint is also employed on KSRC to improve its performance. This is achieved by considering the cooperation among training samples of same class while estimating the SR.

Contents