ADVANCEMENTS IN AUTOMATIC SPEAKER AND SPEECH PROCESSING IN CO-CHANNEL SPEECH

by

Navid Shokouhi

APPROVED BY SUPERVISORY COMMITTEE:

John H. L. Hansen, Chair

Carlos Busso

Issa Panahi

P. K. Rajasakeran

This thesis class file

is dedicated to ...,

who ...

ADVANCEMENTS IN AUTOMATIC
SPEAKER AND SPEECH PROCESSING
IN CO-CHANNEL SPEECH

by

NAVID SHOKOUHI, BS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

November 2016

ACKNOWLEDGMENTS

November 2016

PREFACE

This dissertation was produced in accordance with guidelines which permit the inclusion as part of the dissertation the text of an original paper or papers submitted for publication. The dissertation must still conform to all other requirements explained in the “Guide for the Preparation of Master’s Theses and Doctoral Dissertations at The University of Texas at Dallas.” It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts which provide logical bridges between different manuscripts are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the student’s contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the dissertation attest to the accuracy of this statement.

ADVANCEMENTS IN AUTOMATIC

SPEAKER AND SPEECH PROCESSING

IN CO-CHANNEL SPEECH

Publication No. ____________________

Navid Shokouhi, PhD
The University of Texas at Dallas, 2016

Supervising Professor: John H. L. Hansen

350 word Abstract.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

A wide range of terms have been used to describe various aspects of co-channel speech, which we will to clarify throughout this chapter. We consider both conversational speech and artificially mixed streams as co-channel. All signals treated in this study are single-channel recordings. Of all such data, a subset may have more than one "active" speaker, i.e. multi-speaker speech, which we label as "overlapped speech". Overlapped regions are segments of a co-channel signal where both speakers are active at the same time. This categorization is summarized in Fig. **??**.

The specifics of recording conditions are overlooked in this study. For example, information that relates to each speakers distance from the microphone and room environment. This is intentional, since most of the difficulty in dealing with co-channel speech arises from it being single-channel, which implies our interest in limiting access to spatial information as well as other sources of meta-data.

Alterations in speech production are an important artifact of co-channel speech, which occur solely in conversations and are the result of conscious and/or unconscious reactions of the foreground speaker and interferer(s) during overlaps. Examples of such alterations are raised pitch and energy level [2]. The ESPN show First Take is filled with arguments between the shows two regular sports commentators Stephen A. Smith and Skip Bayless. First Take is a perfect example of an exaggerated version of the above-mentioned changes in speech production. These changes are problematic in automatic speech applications and are considered a type of distortion. Consequently, our treatment will be directed towards applications that suffer the most from such alterations, predominantly speech recognition.

One can argue that co-channel speech has hardly received the attention it deserves compared to other speech related problems. An indication of this remark being the ambiguity in terminology used in the literature. Nevertheless, its presence becomes more noticeable with the increasing demand for automatic speech interfaces as well as the increased diversity where speech data is captured in naturalistic settings. Speech applications have pushed boundaries, forcing researchers to show more interest in tackling problems that have previously been only partially solved.

This dissertation aims to provide tangible solutions to the issue of co-channel speech for automatic speech applications. The first half focuses on speaker recognition/verification by developing an understanding of the effects of speaker interference on various stages of a given recognition system. Assessments are based on different categories of co-channel speech described in the previous section as well as the various standard realizations of a speaker recognition system, namely GMM-UBM (Gaussian mixture model - Universal background model) and i-vector/PLDA (classifying speaker identity vectors, aka i-vectors, using probabilistic linear discriminant analysis). Algorithms developed to improve speaker recognition in co-channel speech can also be used to improve diarization. Most of such approaches attempt to detect regions of overlapped speech. In current speaker diarization systems a significant amount of errors are caused by overlaps [3], hence the importance of overlap detection.

The second half will consider audio stream analysis for diarization with emphasis on speech recognition in co-channel speech. When it comes to speech recognition, aside from the direct effects of interfering speech, the foreground speaker is implicitly affected. This is projected in terms of changes in speech production and mannerisms. It is shown in [1] how an entire word is articulated differently enough that the acoustic model is no longer able to recognize individual phones. This is due to the presence of an interfering speaker, who may not necessarily be active during the time-lapse of the aforementioned word. These are instances in which co-channel speech affects speech production. Changes in speech production have negligible effects on speaker recognition tasks, at least in standard tasks where there is access to considerably large amounts of data to model speaker identities and are therefore overlooked in the first portion of this study. One cannot say the same for speech recognition, since the acoustic models are meant to represent finer acoustic characteristics as compared to speaker recognition. This and many others are among the challenges of speech recognition in co-channel speech.

CHAPTER 2

USAGE INSTRUCTIONS

CHAPTER 3

LITERATURE

3.1 temp

Some text.

CHAPTER 4

CONCLUSION

SAMPLE SOLO APPENDIX

VITA

Navid Shokouhi