SPEECH CHARACTERISTICS DURING CONVERSATION AND THEIR IMPORTANCE ON INFORMATION TRANSMISSION

BOHAN CHEN

Acknowledgement

I would like to express my grate gratitude to my supervisor professor Kazuya Takeda, who has supervised my research at Nagoya University. For the last seven years since I am a master student, he has provided me with a free academic environment in which I could freely explore and pursue my research idea based on my own interests. He has given me many valuable advices for my research, thanks to his broad knowledge and expertise in the field of media signal processing and its applications.

I also would like to express my appreciation to Professor Norihide Kitaoka, who has guided me throughout the years of my studies in Nagoya University. From the initial idea to the final version of this dissertation, he continued to give me invaluable comments and advice to help me. Thank you Kitaoka sensei, thank you for your kind support.

And I would like thank Professor Chiyomi Miyajima and Professor Tomoki Toda, who gave me many valuable comments on my study, and helpful suggestions for my presentations. You are my model of researchers. I feel regret that I can no longer have your comments and advice.

I extend my deep appreciation to all the professors, staff, and my colleagues in the Takeda Laboratory for their continued support, help, and friendship. I also thank all my friends who helped me get through my seven, fantastic years in Japan. Finally, I thank my families, who allowed me to pursue my dream in Japan. Their continuous support encouraged me to be stronger and more confident.

Contents

List of Tables

List of Figures

Abstract

Language production and comprehension have been studied by scientists for decades, And hundreds of empirical observations as well as experimental results have been reported. Numerous models and theories have been proposed to describe a wide ranging aspects of human communication and its underlying mechanisms. This dissertation will investigate the speaker-specific speech characteristics and their impact on information transmission. After reviewing related researches, I consider that similarity in the speaker-specific speech characteristics of conversation partners is one of the key to understanding high linguistic information transmission efficiency during conversations. Since few systematic investigations of the role of speaker-specific speech characteristic similarity in conversational information transmission have been published, especially ones focus on segmental speech characteristics similarity, the objectives of this dissertation are as follows:

1. Investigate the effective similarity measure of speaker-specific speech characteristics on speaker diarization tasks while developing speaker diarization system with higher clustering accuracy (first study);

2. Investigate the impact of similarity in speaker-specific speech characteristics, especially the segmental-based speech characteristics, on subtle prosodic information transmission (second study);

3. Use similarity in the speaker-specific speech characteristics of interlocutors as a predictor of information transmission quality, and explore the effects of this similarity on task performance using spontaneous speech corpus (third study).

In the background survey, segmental-based speech characteristics, prosodic speech characteristics and lexical speech characteristics were found to be the most impor-

tant speech characteristics affecting performance in speaker related signal processing tasks (e.g., speaker identification, speaker recognition, speaker diarization, etc.).

In the first study, a method to evaluate whether a cluster contains all of one (and only one) speaker's speech segments, based on the statistical properties of within-cluster similarity scores and between-cluster similarity scores in a "speaker space" was proposed. Modified bottom-up clustering was then conducted based on the proposed cluster evaluation method in order to increase diarization accuracy by preventing over-merging. Experimental results showed that the proposed method achieve higher clustering accuracy than conventional bottom-up clustering methods.

In the second study, a listening experiment was designed to investigate the efficiency of subtle prosodic information transmission at different levels of speech characteristic similarity. Japanese right-branching (RB) vs. left-branching (LB) ambiguous sentences were used as experimental material. Morphing technology and text-dependent objective similarity measures were introduced to control similarity levels. Participants were asked to finish a target identification task with RB vs. LB materials as targets. Participant response time during the target identification task, as well as the proportion of eye-fixing on different targets were recorded for analysis. Results showed that speech characteristics similarity apparently has facilitative effect on prosodic information transmission.

In the third study, the impact of similarity in speaker and listener speech characteristics on the quality of linguistic information transmission was investigated, using a map task dialogue corpus. Similarity between the segmental (MFCC) voice features, prosodic features and lexical styles of different speakers were analyzed. Most of these similarity measures were shown to have significant facilitative effect on information transmission quality as measured with a direction following task in which a speaker is describing a route to a partner.

In general, experimental results showed that similarity in speaker-specific speech characteristics between conversation partners facilitated information transmission accuracy. The findings imply that self-similar voice is an effective direction towards high efficiency linguistic information transmission. The present dissertation is just the first step towards self-similar high efficiency information transmission system, several important issues, such as the impact of voice naturalness (caused by voice conversion) on linguistic information transmission and the comfortableness of hear-

ing self-similar voice, are still unclear which need to be investigated in the future.

Chapter 1

Introduction

1.1 Overview

As a result of the rapid and continuous advancements occurring in the field of information science, human-like robots are no longer something to be found only in science fiction novels. We now have the technology to dress robots with skin-like materials, teach them how to walk and run on two legs and have them mimic human emotions using various facial expressions. In speech-related fields, systems with the ability to listen and speak to humans have been developed and continue to evolve day by day. A computer's "ears" are driven by speech recognition technology, one of the oldest topics of signal processing research. As a result of technological innovation and several decades of continuous improvement, relatively high accuracy speech recognition systems have been developed, one after another [1] [2] [3]. Current researchers are not satisfied with automated transcription and interpretation of speech content, but also desire to learn more from human speech signals, such as determining the identity of the speaker (e.g., name [4], gender [5], age [6], personality traits [7], etc.), the current status of a speaker (e.g., psychological health [8], stress level [9], intoxication [10], etc.), and information about the speaker's emotional state [11][12]. Using this speech-derived information, speech recognition systems can be customized for specific users, further empowering a computer's "ears" [13] [14] [15] [16].

Similarly, speech synthesis technology, which can provide a computer with a "voice", has also developed greatly in the last decade. Based on our physiological knowledge of speech production and our understanding of mathematical models of random

processes, many facets of synthesized speech are now under the engineer's control [17]. Another product of speech analysis research has been the discovery of speech characteristics which can provide general information about speakers (e.g., vocal spectrogram [18], age [19], etc.), the prosodic characteristics of their speech (e.g., pronunciation [20], sound duration [21], etc.), and their emotion characteristics [22] [23] [24] all of which are available options in current state-of-the-art speech synthesis systems.

In short, modern speech signal processing technology seeks to achieve human-like speech and speech comprehension performance. Computers should be able to not only produce and comprehend linguistic information, but should also be able to respond to nonlinguistic information, such as human identity, physical and emotional states, and behaviour information. Just as it is difficult for humans to interpret linguistic and nonlinguistic information processed separately, knowledge about the integration of this information should help us to design better automated "ears" and "voices".

1.2 Research motivation and contribution

1.2.1 Motivation and objective

In this dissertation, certain speaker-specific speech characteristics containing typical, nonlinguistic information from speech signals are selected, and their importance in linguistic information transmission is examined. Speaker-specific speech characteristics are defined as features which can be used to identify individual speakers. These speech characteristics are typical, nonlinguistic features, but in contrast to other nonlinguistic information, such as information about emotional states, these characteristics of human speech are relatively stable and can be easily extracted from speech signals. Speech can be emotionless, but it is very unlikely to be characterless. However, since a speaker's speech characteristics are determined by various factors, such as their physiological profile, educational background, native language, and so on, there are numerous features which need to be examined. Therefore, my investigations began with speaker diarization, which is a technology that attempts to differentiate and identify speakers without any prior information about them, be-

cause I suspected that speech characteristics which play an important role in speaker recognition may also have an effect on information transmission.

Similarity among speakers is another focus of this dissertation, my hypothesis being that similarity in the speech characteristics of conversation partners plays an important, facilitative role in linguistic information transmission during dialogue. As I will discuss in the following chapter, it is believed that the language communication process is an attempt to reconstruct a situation model orally and mentally. Similarity in the speech characteristics of speakers are considered to facilitate this process directly in two ways. First, it can increase linguistic information retrieval efficiency due to listener familiarity with the speaker's manner of speaking, freeing up cognitive resources for listener comprehension. Second, it can increase utterance prediction accuracy among interlocutors. Although it seems obvious that familiar (self-similar) accents and vocabulary are easier to comprehend and produce, there are issues which need to be examined. For example, the relationship between similarity in segmental speech characteristics (e.g., MFCC), one of the most important types of speech characteristics used in both laboratory and commercial research, and information transmission, has never been thoroughly investigated. Although it is believed that segmental speaker characteristics mainly contain air vibration information, which is different from the bone conducted voice we hear when we ourselves speak, it is likely that segmental characteristics also contain information (e.g., dynamic spectrum information) which we may be familiar with, as a result of the thousands of conducted and overheard conversations we have experienced. Therefore, I consider similarity in segmental speech characteristics as potential factors which can affect the information transmission process, and have decided to investigate their influence in this dissertation. Additionally, previous studies have conventionally used categorical similarity levels (e.g., local accent versus foreign accent) to examine the impact of speech characteristic similarity on information transmission, but more quantifiable measures of speech characteristic similarity are necessary for further investigation. Finally, most previous studies have utilized laboratory experiments conducted under strictly controlled conditions, thus there is little data on the impact of speech characteristic similarity on information transmission in spontaneous dialogue corpora, which is obviously nearer to real world conversation conditions.

1.2.2 Contribution

The aim of this dissertation is to investigate which speaker-specific speech characteristics influence information transmission efficiency. I will mainly focus on the role of similarity in the speech characteristics of conversation partners in order to gain a deeper comprehension of the mechanisms of the human language process. For this reason, speech diarization was first studied in order to understand the importance of various speech characteristics on speaker recognition and the possible approaches to measuring the similarity of speech characteristics. A novel clustering method was also proposed for high-accuracy speaker diarization. Experimental results showed that the proposed method increased the accuracy of speaker clustering by the bottom-up clustering algorithm.

Several methods of measuring the similarity of important speech characteristics, including segmental characteristics, prosodic characteristics and lexical characteristics were then selected, and their impact on subtle prosodic information transmission efficiency was tested using a behavioural experiment. Analysis of the response time data showed that participants understood prosodic information more quickly when it was communicated by voices similar to their own. Analysis of the visual fixation data also showed that participants understood more of the prosodically conveyed information when the target images were described in voices similar to their own. These findings were consistent with one another, and imply that acoustic feature similarity is relevant to prosodic information transmission efficiency. Moreover, based on an analysis of the objective, text-dependent similarity measurements employed, differences in prosodic expression between paired participants were subtle. As it is unlikely that these subtle prosodic differences have any linguistic meaning, our results suggest that human information processing is so sensitive that even subtle speech characteristics of speakers can influence information processing efficiency, as is also the case regarding spectrum similarity (MFCC distance), which is considered to contain information on the condition of the vocal tract.

Finally, I wanted to expand my inquiry into the relationship between speech characteristic similarity and linguistic information transmission quality from laboratory experiments to spontaneous speech. To do this, similarity in the speech characteristics of speakers was proposed as a predictor of information transmission quality in dialogues from a spontaneous speech corpus collected during a direction/navigation

task. Using approximated measurements of linguistic information transmission quality, the effectiveness of each predictor were analyzed at different levels, a topic which had previously been little researched. Experimental results showed that most of the speech similarity measures selected had significant facilitative effect on linguistic information transmission accuracy. The results of this investigation suggest that synthesized, listener-like speech could be used for high accuracy information transmission.

1.3 Outline of the dissertation

Outline of the dissertation is shown in Fig 1.1.

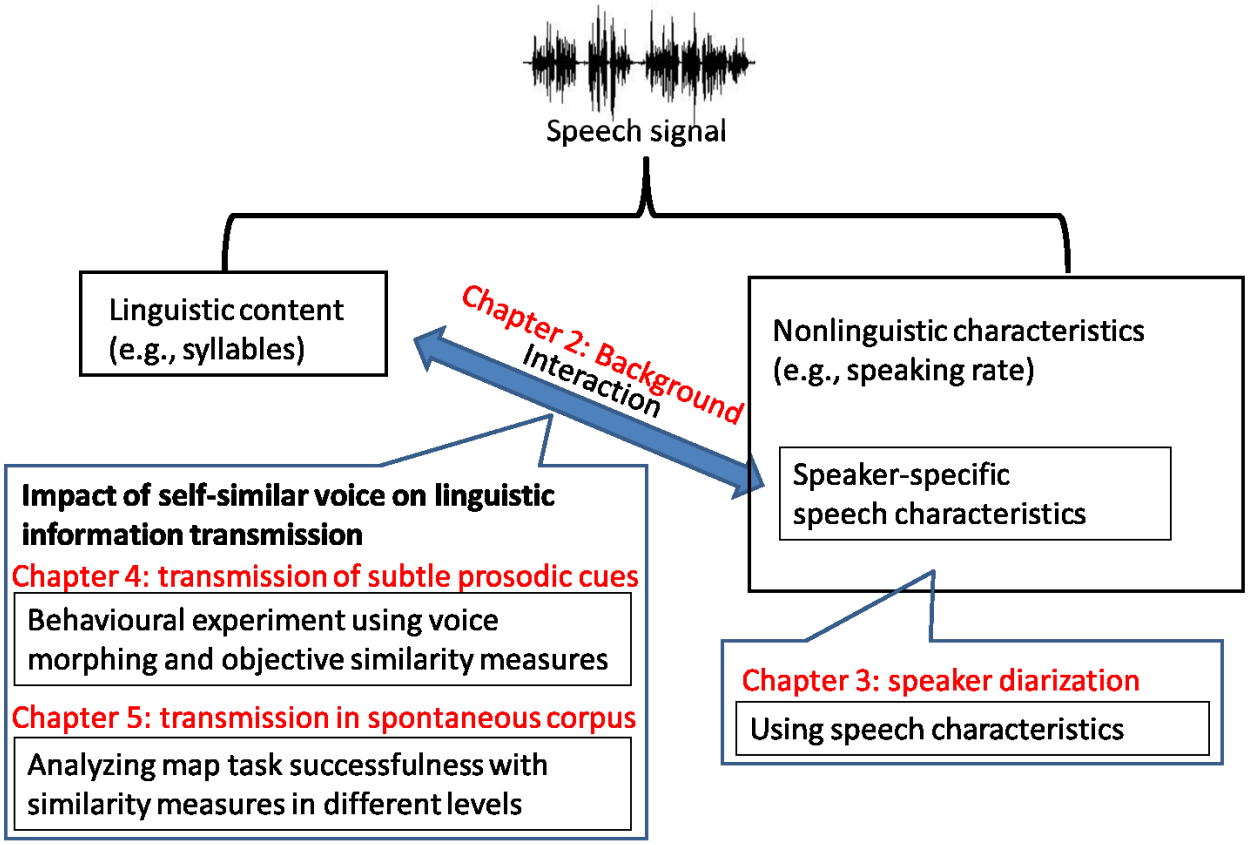

Figure 1.1: **Outline of this dissertation.**

In Chapter 2, I provided a brief review of the literature on speech characteristics and language processing, as well as of studies on the relationship between them. After

describing the development of theories in these fields of research, I then proposed the main hypothesis of this dissertation.

In Chapter 3, I proposed a method to increase clustering accuracy in connection with a speaker diarization task. The proposed method is then evaluated using an open-source spontaneous meeting speech corpus.

In Chapter 4, a behavioural experiment designed for investigating the impact of similarities in speaker characteristics on the efficiency of subtle prosodic information transmission is described.

In Chapter 5, I tested the idea that similarity in speech characteristics can be used as a predictor of information transmission quality in dialogues. Spontaneous speech data was used to test my hypothesis, and the effects of similarity in various speech characteristics on information transmission quality were investigated.

Finally, in Chapter 6, I drawn my conclusions and discuss possible future work.

Chapter 2

General Background

Keywords in this chapter:

- Speech characteristics
- Speaker recognition
- Information transmission in dialogue
- Psycholinguistic theory
- Conversational alignment
- Similarity of speech characteristics

2.1 speech characteristics

Speech signals convey a mixture of information including linguistic information and speaker- specific information. Speakers of different ages, genders, education levels, social classes, etc., tend to speak in different ways based on their identity. For example, imagine a male, Chinese student giving a speech in English at an international conference. His voice may have a lower pitch because men's vocal folds tend to be larger than women's. His voice may waver because he is a student who is not accustomed to public speaking, and his nervousness may also affect the fundamental frequency and spectrum of the sounds produced at his glottal source. Since he is Chinese, the vocabulary he choose, his pronunciation and his speaking rate may all be different from those of a native English speaker, or even from another non-native speaker from a different nation. All of these features which can be obtained from a speech signal can potentially give us the ability to discriminate one speaker from others, and these features are what are referred to as speaker-specific speech characteristics in this dissertation.

There are several possible ways to categorize speech characteristics. Based on their continuity, they can be categorized as discrete characteristics, which applies to most of the linguistic characteristics, such as one's vocabulary, or continuous characteristics, such as the pitch contour of one's speech. Based on their stability, they can be divided by congenital characteristics which are unlikely to change (e.g., glottal source frequency), acquired characteristics which are likely to remain unchanged during a specified period (e.g., pronunciation), and temporal characteristics which are likely to change often based on the speaker's mental condition and the surrounding environment (e.g., speech intensity). Or, based on the size of analysis windows, speech characteristics can be separated into short-term characteristics (e.g., pitch, which is extracted every 20ms in general) and long-term characteristics (e.g., average pitch of a speech segment or a speaker). In this dissertation, speech characteristics will be categorized based on their temporal spans as follows: 1) segmental speech characteristics, 2) prosodic speech characteristics, and 3) idiolect speech characteristics, each of which will be defined in the following subsections of this dissertation.

In general, segmental speech characteristics are easy to compute and yield good performance for both speech and speaker recognition [25]. Prosodic speech characteristics are considered to have higher robustness to channel variation (i.e., noise coming from the acoustic environment and other technical factors), but are less useful for discriminative tasks and are easier to imitate, especially by professional voice imitators [26]. Idiolect speech characteristics, while sharing many of the strong and weak points of prosodic speech characteristics, usually require the use of preprocessing support (e.g., an automatic speech recognizer) with an acceptably high accuracy, which is still difficult to achieve. Therefore, there is no best type of speech characteristics for general use, so the choice of which speech characteristics to use should be based on the particular application they will be used for. In general, a combination of different speech characteristics will likely be the most useful for discriminating among and identifying different speakers [27] [28].

2.1.1 Segmental speech characteristics

Here, segmental speech characteristics refer to characteristics that calculated from the smallest discrete linguistic unit that can be identified. In practice, segmental speech characteristics particularly refer to information that extracted from a

short-term analysis window(s) (usually $20ms$ $30ms$) which is considered to contain phonemic information.

One of the most important segmental speaker characteristics is the spectrum of the speech signal. As the sound of the speech we hear is actually the result of the vibration of air molecules under the effect of sound waves. Because information about our speech organs (e.g., our vocal tract) is contained in the air vibrations they produce, the frequency domain information of these vibrations can provide not only linguistic information but also information about speaker-specific speech characteristics. The first step towards spectrum analysis of speech characteristics is usually the transformation of the signal from a time domain signal to a frequency domain signal. As a speech signal continuously changes due to the articulatory movements of the vocal tract, transformation usually occurs in short time frames with an interval of $20-30ms$. The signal is assumed to remain stationary within this interval, and a spectral feature vector is then extracted from each frame. Because of the finite-length effects of the discrete Fourier transform (DFT), the frame is multiplied by a window smoothing function, usually a Hamming window, before further processing.

There are several ways to quantify a spectrum, but use of mel-frequency cepstral coefficients (MFCCs) [29] is one of the most popular methods in speech and audio processing. The calculation of MFCCs generally has three steps:

1. Apply to the spectrum a bank of triangularly shaped filters whose centers are evenly spread over the mel-frequency axis;
2. Calculate the natural logarithm;
3. Apply a discrete cosine transform (DCT).

MFCCs use mel-frequency scaling instead of frequency scaling to create sub-band filters. The human acoustic system has different resolution ratios in different frequencies, and filter banks based on mel-frequency scaling are able to catch more of the low frequency information produced by human beings [30]. Using a source-filter model, speech spectrum $Y(\omega)$ is considered to be the product of two components which are assumed to be independent of each other: 1) a source component $X(\omega)$, which mainly represents the vibration of the vocal folds in response to airflow from

the lungs (voiced speech) or turbulent airflow in the vocal tract (unvoiced speech); and 2) a filter component $H(\omega)$, which mainly represents the effects of the vocal tract and lip radiation [31]. The high frequency region of the spectrum is believed to contain information about the source component, while the low frequency region of the spectrum is believed to contain the filter component. Using the calculated logarithm with the inverse Fourier transform (using DCT) is a mathematical procedure which tends to separate the two components.

Although using MFCCs is the most popular spectrum speech characteristic analysis method, there are several alternative methods which can also be used to capture speech characteristic information from a spectrum. Linear predictive cepstral coefficients (LPCCs), for example, use the cepstrum (i.e., inverse Fourier transform of the log-spectrum) of the linear prediction coefficients (LPCs) as speech characteristics [32]; perceptual linear prediction (PLP) coefficients utilize critical-band spectral resolution, equal-loudness curves and the intensity-loudness power law as a psycho-physics basis for computing speech characteristics[33]. It has been noted that most spectrum features, including MFCCs, were originally proposed for use in automatic speech recognition (ASR), which means that these features were originally purposed to extract more linguistic information while ignoring the speaker-specific information. In contrast, spectrum features that can catch more speaker-specific information while ignoring linguistic information has also been designed[34].

Furthermore, the spectrum speech characteristics described above are extracted from a single frame. As speech signals change continuously, it is reasonable to assume that spectro-temporal signal details, such as formant transitions and energy modulations, contain useful speaker-specific information which can be extracted as speech characteristics. Although researchers have used spectro-dynamic speech characteristics in attempts to identify robust, discriminative speech characteristics [35][36], simple differentiation using the time deviation of features (e.g., MFCCs) seems to yield equal or even better performance in practice [37]. Simple differentiation is usually computed using time differences between adjacent speech characteristics at the frame level [38]. The first order time deviation is called delta (Δ), and the second order time deviation is called delta-delta ($\Delta\Delta$). The original speech characteristics are mathematically expanded three times (i.e., MFCCs with 20 dimensions will be expanded to 60. dimensions when the delta and delta-delta values are also included).

Another set of segmental speech characteristics is pitch and intensity related features of the speech signal, which calculated based on the pitch and intensity values computed within a short-term time frame with an interval of about 30 ms . They are basic elements of sound wave, and are considered to be speaker-specific speech characteristics because they can reflect, 1) physiological characteristics of speech production organs and 2) acquired/learned habits. For instance, the amplitude variation is affected by the reduction of glottal resistance on the vocal cords and is correlated with the changes in noise emission and breathiness. Researchers have investigated methods to measure this variation (known as "shimmer") and have used it for speaker recognition [39]. In [40] researchers investigated seventy different such features and their speaker identification performance when combined with conventional MFCCs in a speaker diarization task. Pitch based features showed the best performance.

Pitch, also known as fundamental frequency (F_0), is one of the the most important speech characteristic[1], and is considered to reflect the speaker's vocal fold vibration rate. It is affected by the various physical properties of the speaker's vocal folds, including their size, mass and stiffness, which obviously contain a large amount of speaker-specific information [41]. For instance, using the parameter of mean pitch alone can provide language-independent gender identification with 98% accuracy [42].

There are also speech characteristics developed based on the frequency and amplitude properties of sound wave, such as jitter and shimmer, which are two of the so-called "voice quality" features. Jitter is defined as pitch variation from cycle to cycle, and shimmer is defined as the variation in the amplitude of a sound wave[2]. Since jitter and shimmer characterize aspects of particular voices, it is a priori expected that differences in values for jitter and shimmer will be found among different speakers. The usefulness of jitter and shimmer for speaker identification has been tested in several studies which have shown that, as with other prosodic speech characteristics, jitter and shimmer can improve speaker recognition accuracy

[1]Strictly speaking, pitch is the perceived frequency of a sound, which determined by comparing a signal with pure sinusoids to determine the best match, and is thus a psycho-acoustic concept. Fundamental frequency, on the other hand, is the lowest frequency of a periodic waveform, and is thus a physical concept. I will use the two terms interchangeably in this dissertation, however.

[2]Several methods of computing these features have been proposed. For example, the famous speech analysis toolbox Praat[43] contains five different computations to determine jitter "frequency variation" ("local", "local, absolute", "rap", "ppq5", and "ddp").

when combined with conventional spectral features such as MFCCs [39][44]

2.1.2 Prosodic speech characteristics

Prosodic speech characteristics are usually known as the suprasegmental speech characteristics, which refer to speech characteristics that extend over syllables and longer regions. Figure 2.1 shows an example of the usefulness of prosodic speech characteristics (i.e., pitch contours) for speaker identification, from [45]. Four speakers (two males, one female, and a child) repeated the word "sunday" three times, and the pitch contours of their speech are shown in Figure 2.1. We can see that the pitch contours are similar when the word is spoken by the same speaker, but quite different when the speaker changes, even though the subjects all uttered the same word.

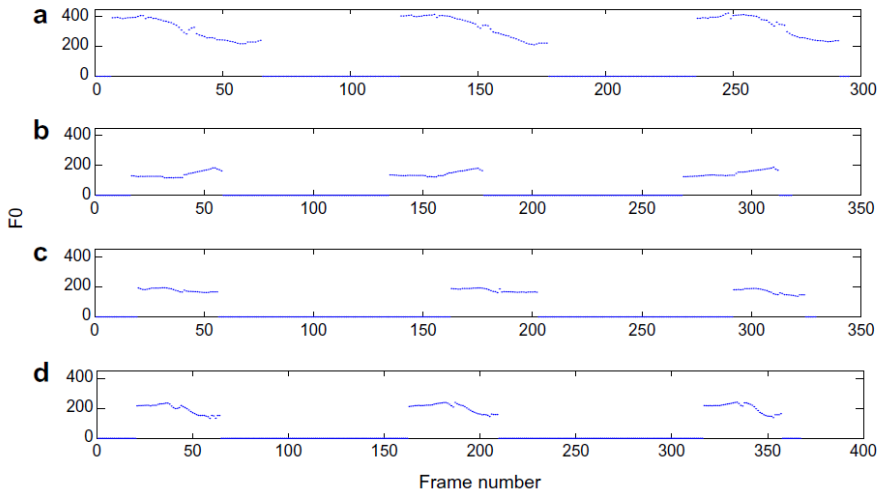

Figure 2.1: **Variation in F_0 contour dynamics of four different speakers: : (a) Child, (b) Male 1, (c) Male 2, (d) Female. All four subjects repeatted the same word three times ("Sunday", "Sunday", "Sunday") [45].**

Researchers have proposed several methods of modeling prosodic features for long-term intervals to determine prosodic speech characteristics. One straight-forward method is described in [46]. The tracked pitch contour of a syllable-like segment was modeled by fitting it with a piecewise linear model, and the parameters of the linear model were then used as statistical prosodic speech characteristics. In [47], parameters were simplified into binary or integer values, and symbolic sequences of

the coded prosodic features (i.e., rising vs. falling pitch contour, rising vs. falling intensity contour, short/median/long duration) were then used as prosodic speech characteristics. Other prosodic speech characteristics which have been extracted from piecewise linear models of pitch contours can be found in [48][49][50].

2.1.3 Idiolect speech characteristics

An idiolect is the language or speech of an individual during a particular period of his or her life. Here idiolect is used to refer to speech characteristics extracted from morpheme (i.e., the smallest meaningful unit) and longer regions. In contrast to prosodic speech characteristics, idiolect speech characteristics are all linguistically meaningful. Idiolect speech characteristics can be useful for speaker identification in two ways. First, when combined with other features, idiolect speech information can be used to model the physiological characteristics and speech habits of a specific speaker. Second, idiolect speech information can be used as speech characteristics themselves, alone or together with other speech characteristics. For example, in [51] Doddington used the relative frequency of common words to identify speakers.

Before obtaining idiolect speech characteristics, it is always necessary to separate the original speech signal into segments with meaningful linguistic information, either manually or automatically. These segments, also known as "tokens", include words [51], and even part-of-speech (POS) tags [52]. Note that when we consider words or even longer idiolect features as speech characteristics, the techniques which are used to identify speakers can also be used to identify writers[3].

Segmental and prosodic speech characteristics can be modeled not only by using whole speech, but also by using information within tokens (e.g., words). In [53], researchers provided a method of modeling phonetic MFCCs using Gaussian mixture models (GMMs). Based on an automatic speech recognition (ASR) system, this method trained speaker-specific GMMs for every phone unit for speaker recognition. Boakye and Peskin [54] proposed a similar method, using an ASR system to provide hidden Markov models (HMMs) with a set of high-frequency keywords used by a specified speaker as speech characteristics. Experimental results show that both of these approaches can be used to obtain speech characteristics which can be used for

[3]In scientific literature, the history of authorship recognition is longer than that of speaker recognition.

state-of-the-art speaker identification.

When using only idiolect tokens for speaker recognition, it is common to use the N-gram probability of token tags as speech characteristics. N-gram models can determine the conditional probability of a token sequence. For example, $N = 2$ results in a bigram which calculates the appearance probability of $t_i + 1|t_i$, where t_i is the i-th token tag of a specific speaker's speech. Word N-gram model have been widely used for both speaker recognition [51] and authorship recognition [55] (which can also be viewed as speaker recognition based on manual transcription), but the accuracy level is comparable to methods using individual words [56]. Other researchers, with the help of natural language processing (NLP) techniques, have used N-grams to model syntactic information or other linguistic speech features as speech characteristics [57]. Idiolect speech information can also be used alone. The appearance frequency of a set of keywords is the usual method when using this approach. Function words such as "for", "the", "and", etc. are usually chosen as keywords for three reasons: 1) function words are content/topic-independent, and thus can be used as a general measure; 2) function words are generally commonly used words, thus it is possible to collect sufficient data from a limited amount of speech; and perhaps most importantly, 3) function words are considered to be used unconsciously and are therefore important markers of a variety of individual differences and social behaviors, ranging from leadership style to honesty [58][59].

2.2 Information transmission during dialogues

Information transmission can be described as information production by a sender, the transfer of that information through some medium, information reception by a receiver, and comprehension of that information by the receiver. As humans mainly engage in linguistic information transmission, the issue can be simplified as language production (by a speaker or writer) and comprehension (by a listener or reader), topics which have been widely researched in the last three decades. This section will provide a brief review of psycho-linguistic research, which will also provide an explanation of concepts related to language production and comprehension. Although traditional theorists consider language production and comprehension as two independent processes, recent empirical investigations from behavioral and neuro-

imaging studies show that they tend to be intertwined [60][61]. As a result, this section will not discuss these two issues separately.

Most psycho-linguistic theorists, including both historical and modern researchers, consider the process of linguistic communication as action that begins with the linking of linguistic units (e.g., words, clauses) with their referential units (e.g., actions, events) in the mind, followed by an effort to integrate them into an oral/mental representation [62][63]. It is also understood that short-term and long-term memory play an important role in the language production process. Short-term memory not only stores the previous linguistic units of the current utterance, but also information about previous utterances in the series, which affects the mapping and interpretation process both consciously and unconsciously. Meanwhile, parts of the long-term memory are used for this process by providing schemata, experience and linguistic knowledge[64].

Linguistic priming is a good example of how information stored in short-term memory affects language processing. Based on psycho-linguistic studies which have focused on the priming effect, information that is stored in our short-term memory affects both our language production and language comprehension processes. Like a prompt, humans tend to produce information in a style similar to the information stored in their short-term memory, and to also comprehend what they hear based on the style of the information stored in their short-term memory. Here, "style" includes almost all linguistic concepts, from the phonetic to the semantic level. An example of lexical priming is provided by [65], in which researchers showed that people were faster at recognizing a string of letters in a word when the word followed a semantically related word (e.g., "nurse" is recognized more quickly when it follows "doctor" than when it follows "bread"). An example of syntactic (or structural) priming can be found in another early study [66], in which researchers observed that a strong predictor of the use of the passive tense in interviews was the presence of another passive sentence in the previous five utterances. Another more recent study on syntactic priming [67] involved an experiment in which participants first saw an ambiguous sentence, such as:

- The policeman prodding the doctor with the gun,

which can be understood as either "policeman used the gun to prod the doctor", or "the policeman prodded the doctor who had the gun". The subjects were then shown

two pictures, one of which matched one or the other of the possible interpretations, and another which matched neither interpretation. Participants were then presented with a syntactically similar target sentence, such as :

- The waitress prodding the clown with the umbrella,

and were then shown two pictures, but this time each picture matched one of the possible interpretations of the target sentence. Researchers found that the participants tended to choose the picture that matched the syntactic structure of the sentence in the first trial which matched the picture in the first trial. Other studies [66]have also suggested that some of our analysis and production choices are affected by the information stored in our short-term memory. For more examples and discussions about priming, see review paper [60]. This priming effect is so common that it has been observed in a broad range of studies, in situations from monologues to spontaneous conversations; almost everyone's language comprehension and production styles are affected by what they have just experienced.

Schema theory is a typical theory used to explain the influence of long-term memory. In one of the field's original studies, a researcher asked British subjects to reproduce a native American legend they were told, and found that participants tended to replace some details involving things they were unfamiliar with with things they were more familiar with (e.g., they replaced "peanuts" with "nuts") [68]. The author of the study hypothesized that language comprehension is an active process, and that received information only provides a general outline for listeners, who then retrieve or construct a mental representation based on their prior knowledge structures (schemata)[69]. Some modern experiential theorists have expanded this view and claim that the listener is immersed in a mental representation which they themselves have constructed [63]. For example, the sentence "The mouse approached the fence" implies not only a different distance between the subject and object of the sentence than does "The tractor approached the fence", it also implies a different distance between the observer and the situation being observed [70]. Empirical evidence has been provided by both behavioral and neuro-imaging studies. For example, it has been found that words activate areas of the brain that overlap with areas that will be active when their referent is experienced. Nouns which are difficult to visualize (e.g., "justice") are less likely to activate visual areas of the brain than concrete nouns (e.g., "apple") [71]. Behavioral studies have also shown similar results, for

instance, that subjects respond more quickly to a picture of an eagle with its wings spread out after reading "The ranger saw an eagle in the sky" than after reading "The ranger saw an eagle in a tree" [72]. The authors claim that due to the subjects' experience of viewing the picture of the eagle with spread wings, the location of the eagle is constrained by its experienced appearance.

Meanwhile, long-term memory also facilitates language comprehension as a predictor [61]. When participants read sentences such as "The day was breezy so the boy went outside to ...", researchers observed an N400[4] effect when the sentence ended with the less predictable "an airplane" than with the more predictable "a kite". Curiously, the N400 effect occurred at "an", not at "airplane". Researchers argued that the N400 effect was the result of the subjects' discovery that they had made a mistaken prediction based on the appearance of the word "fly", which they realized as soon as the phonological form of the expected word (i.e., that it began with a consonant) was eliminated by the appearance of the word "an" [73]. A similar N400 effect was also observed while participants heard sentences such as "I like to drink wine" spoken by a child instead of by an adult voice [79], which provides evidence of prediction based on social commonsense.

2.3 Impact of speech characteristics on information transmission in dialogue

The first question to be addressed in this section is whether or not the speech characteristics of a speaker have an impact on information transmission in dialogues. Traditionally, researchers have tended to answer "yes", because they have found that participants can remember fewer words when word lists are presented by multiple speakers than by a single speaker, as in [75] for example. However, in [76]researchers found the opposite, i.e., that word lists spoken by multiple speakers can be more accurately remembered if the duration of the pause between words is increased (e.g., from 250 ms to 4000 ms). Researchers have also argued that speaker recognition and speech recognition are processed by different parts of the brain. Aphasia research has shown that patients with damage to the right hemisphere of the brain

[4]N400 is a negative electroencephalography (EEG) signal observed when subjects encounter unexpected words in a read sentence.

perform worse in voice discrimination tasks [77]. Moreover, recent neuro-imaging studies have found consistent evidence that there appears to be more activity in the right hemisphere than in the left hemisphere when subjects recognize a speaker's voice [78]. Therefore, the answer to the question of whether or not speech characteristics have an impact on information transmission in dialogues still tends to be "yes". But it is also necessary to note that, as speech characteristics and linguistic information are probably processed in two separated areas (i.e., in different hemispheres) of the brain, the impact of speech characteristics may not be very strong.

The second question which needs to be addressed is how the impact of speech characteristics on language comprehension is manifest. As mentioned above, short-term and long-term memory are two important factors in the human language process, so in the following sections how speech characteristics stored in short-term and long-term memory affect our transmission of information in dialogues will be examined. For the impact of short-term memory, I will mainly introduce researches focus on speech characteristics alignment. For the impact of long-term memory, in the other hand, I will mainly introduce researches focus on the familiarity of speech characteristics.

2.3.1 Impact of speech characteristics alignment in dialogue

Since the priming effect automatically and ubiquitously affects the communication process, it is not hard to imagine that one of its natural results is the linguistic alignment of speakers during dialogues. Because speech stored in short-term memory has a higher reproduction probability (or even repetition probability, in some linguistic aspects), it is likely that as a conversation goes on, the interlocutors' speech patterns will align with each other. In [74], the authors summarized the alignment phenomenon and proposed the idea of interactive alignment, which is the hypothesis that alignment occurring at one linguistic level will probably lead to alignment at other linguistic levels, and when interlocutors achieve mental states which are in alignment (i.e., when they understand a topic in a similar way), conversations tend to be more successful.

Numerous studies have investigated alignment in speech characteristics during di-

alogue. Some of these studies focused on whether or not alignment occurs and exactly which speech characteristics are affected, while others attempt to quantize that alignment or have proposed scales of conversational successfulness. In [80] for example, researchers investigated the alignment of prosodic speech characteristics such as mean pitch using a cooperative game corpus, and found that prosodic speech characteristics did indeed become more similar during conversation. Lexical alignment, the choice of backchannel words such as "well" during a conversation, was also investigated in [81]. The results of that study showed that similarity in the use of backchannel words increases during conversation. Researchers have also had positive results when using alignment as a measure of dialogue successfulness. For instance in [82], researchers found that alignment in the use of high frequency words had a positive correlation with both the successfulness and naturalness of the dialogue. Similar results have also been found when researchers have investigated the relationship between dialogue successfulness and syntactic alignment using another cooperative conversation corpus[83].

2.3.2 Impact of familiarity of speech characteristics on information transmission during dialogues

As mentioned in Section 2.2, familiar (self-similar) speech characteristics can affect information transmission in two ways. First, since linguistic units (e.g., words) are stored structurally, their appearance with familiar speech characteristics tends to help us retrieve their schemata more easily. For example, the opposite results of studies [75] and [76], which I described at the beginning of this section, can be explained if we assume that words are encoded along with the speaker's speech characteristics. When participants do not have enough time for encoding (i.e., in [75]), they remember less information because part of their mental processing capacity is needed to adjust to the speakers unfamiliar speech characteristics. When participants do have enough time (i.e., in [76]), speech characteristics help listeners retrieve the fully encoded words, and therefore they are able to achieve better recall performance.

Second, because interlocutors with similar speech characteristics also tend to be similar in physical, social, cultural and educational backgrounds, these similarities will increase the predictability of their representations and therefore positively impact

information transmission[61]. Moreover, interlocutors tend to show more friendliness towards people who have speech characteristics they are familiar with. This suggests that interlocutors sharing similar speech characteristics will put more effort into cooperation, which might indirectly lead to a more successful dialogue. In [84] for example, researchers found that infants preferred to look at faces associated with their native language, and that young children preferred to be friends with speakers of their native language or speakers who had native accents. Furthermore, in [85] researchers found that people with similar styles of language use tended to develop relationships with each other, and were able to maintain those relationships longer.

2.4 Summary of this chapter

In Section 2.1, various features which have been developed for speaker recognition were reviewed. Although segmental features are the most widely used and most effective features in both laboratory and commercial settings, other features such as prosodic and lexical features, which are considered to contain additional speaker identification information, are also useful. In Section 2.2, modern linguistic theories about the language process were reviewed. Most theorists accept that interaction between short-term memory, long-term memory and input information play a crucial role in language production and comprehension. Short-term memory not only provides contextual information but also affects our interpretation and production preferences at all linguistic levels. Meanwhile, linguistic units (e.g., words) are stored together with memories of our personal experiences in our long-term memory. Hence, retrieving a linguistic unit will probably activate the related experience, and vice versa. In Section 2.3, studies investigating the relationship between speech characteristics and information transmission in dialogues were reviewed. Many studies on conversational alignment have been conducted, and although different methods of quantizing linguistic alignment have been proposed, this alignment has been successfully used as a predictor of dialogue success. However, there have been few studies which have systematically investigated the relationship between similarity in speech characteristics and success in information transmission. In addition, in contrast to the important role that segmental speech characteristics plays in the field of speaker recognition, the correlation between segmental speech characteristics and information transmission efficiency have not apparently been investigated yet. This may be

natural when we consider that segmental speech characteristics are hard to imitate, making them useful for speaking identification, but when we consider the correlation between similarity in speech characteristics and information transmission efficiency in dialogues, it is worthwhile to investigate not only the role of prosodic and idiolect speech characteristics, but also that of segmental speech characteristics, which is one of the motivations of this dissertation.

Chapter 3

Modified Bottom-up Clustering Based on Cluster Evaluation and Removal of Distinctive Speaker Data for Improved Speaker Diarization

Keywords in this chapter:

- Speaker diarization
- Cluster evaluation
- Modified bottom-up clustering

In this chapter, I study on a straight-forward application of speaker-specific speech characteristics, speaker diarization, which is treated as a speech segment clustering problem based on speaker-specific speech characteristics. I propose a method to judge whether or not a speech cluster is composed almost entirely of speech segments from only one speaker, and adopt this method for bottom-up clustering for speaker diarization. Such clusters are ideal for speaker diarization, and detecting and filtering these clusters during bottom-up clustering can make clustering more effective and efficient. The procedure utilizes statistical characteristics of the distance between the centroid of a cluster and each initial speech segment. Experimental results show that the proposed method can effectively distinguish at least one speaker's data from all other speakers in multiple audio files. When the proposed cluster eval-

uation method was applied to our modified bottom-up clustering algorithm, higher clustering accuracy was achieved compared to conventional bottom-up baselines.

3.1 Introduction

Speaker diarization is a task that attempts to answer the question "who spoke when" without any prior knowledge about a conversation (e.g., a meeting, broadcast, etc.). It is usually treated as a speech segment clustering problem, in which the speech of each speaker is clustered into discrete groups [86]. The ideal result of speaker diarization is that each cluster is composed almost entirely of speech segments from an unique speaker. The conventional approach to speaker diarization involves three steps: 1) during pre-processing, non-speech sections identified using voice activity detection and sections with overlapping speech are detected and removed; 2) during segmentation, the remained speech data is divided into initial segments, each of which are assumed to contain the speech of only one speaker; 3) during clustering, speech segments are clustered based on the speech characteristics of the speaker [86]. Segmentation can begin with initially rough segments (e.g., speech is randomly divided into N segments) and then re-segmented during clustering [87]. An alternative method is to use a relatively small segment duration (e.g., one second), which is likely to contain only one speaker's speech [88]. As segmentation does not usually affect the final diarization results too much, clustering is considered to be the key component of speaker diarization systems, and is therefore the main issue of this study.

Having no prior knowledge about the nature of the conversation or the characteristics of the speakers is one of the reasons why speaker diarization is such a difficult task. Both the number of speakers and their speech characteristics need to be estimated based only on the target speech[1]. Conventional diarization methods tend to treat every speaker in a conversation equally, which means that when clustering stops, every cluster which remains is hoped to be an ideal cluster composed only of the speech of one speaker [87][89]. However, the speech characteristics of various speakers vary in their speech characteristics similarity. Some speakers are easier to discriminate from others. Imagine doing speaker diarization manually, the separation of speakers

[1]For other speech characteristic related tasks, such as speaker recognition, the speech characteristics of the speakers can usually be obtained in advance from training data.

who have distinctive voices initially should obviously increase diarization accuracy. Hence, I assume that the accuracy of automated speaker diarization can also be improved using a similar approach.

In this chapter, I proposed a method to evaluate whether or not a cluster is an ideal cluster (i.e., whether or not it is composed almost entirely of speech segments from an unique speaker). A modified bottom-up clustering algorithm is then proposed based on this cluster evaluation method. Experimental results show that proposed method increases diarization accuracy.

The rest of this chapter is organized as follows. After a description of the proposed cluster evaluation method in Section 3.2, the modified bottom-up clustering algorithm based on the proposed evaluation method is described in Section 3.3. In Section 3.4, the results of the clustering experiment are reported. I end the chapter with my conclusions and a discussion of future research.

3.2 Cluster evaluation based on a speaker space

In this section I propose a method to evaluate whether or not a cluster is an ideal cluster containing all of the speech segments of a specified speaker ($recall \approx 1$) and only containing speech segments from that particular speaker ($precision \approx 1$). I assume that in a speaker space where speech characteristics are well represented, clusters which approximate an ideal cluster shows some special statistical properties between intra-cluster segments and inter-cluster segments, which will be used for cluster evaluation.

3.2.1 Speaker space

In addition to speaker-specific speech characteristics, audio signals contain other information which would influence the accuracy of analysis if we do nothing about it. In order to extract speaker-specific speech characteristics from audio signals, researchers have proposed various vector spaces which are also known as speaker spaces [90][91]. The speaker space proposed in [90] (a so-called i-vector) was used in this study. This method uses factor analysis to extract vectors from an audio signal to represent speaker-specific speech characteristics[90][92]. A speaker- and

channel- (or session-) dependent supervector $\mathbf{M}$, which is usually composed of the mean vectors of the Gaussians in the Gaussian Mixed Model (GMM) trained from the current data set, can be expressed as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{3.1}$$

where $\mathbf{m}$ is the speaker- and channel-independent supervector, commonly taken from a universal background GMM representing speaker- and channel-independent acoustic features. $\mathbf{T}$ is a rectangular matrix using virtual speakers as its rows (i.e., it is the speaker space), and $\mathbf{w}$ is a vector called the total factor vector or i-vector.

Cosine similarity has been applied successfully in the speaker space to measure the similarity of two i-vectors. Therefore, the similarity of two data sets, X and Y, represented by two vectors in the Total Variability space (the speaker space) $\mathbf{w}_X$ and $\mathbf{w}_Y$ respectively, can be measured using the following equation:

$$\sigma(\mathbf{w}_X, \mathbf{w}_Y) = \frac{(\mathbf{w}_X)^t(\mathbf{w}_Y)}{||\mathbf{w}_X|| \cdot ||\mathbf{w}_Y||}. \tag{3.2}$$

3.2.2 Cluster evaluation

Based on the speaker space introduced in Section 3.2.1, in this section I propose a method to evaluate whether or not a cluster is a ideal cluster. Here the i-vector of cluster α in speaker space was noted as $\mathbf{w}_\alpha$[2]. To evaluate $\mathbf{w}_\alpha$, I define $\mathbf{D}_\alpha$ as the set which contains the similarity data between $\mathbf{w}_\alpha$ and all the segment vectors in the speech. Therefore $\mathbf{D}_\alpha$ can be written as follows:

$$\mathbf{D}_\alpha = \{\sigma(\mathbf{w}_\alpha, \mathbf{s}_1), \sigma(\mathbf{w}_\alpha, \mathbf{s}_2), ..., \sigma(\mathbf{w}_\alpha, \mathbf{s}_N)\}, \tag{3.3}$$

where $\mathbf{s}_i$ represents the i-vector of segment i, N is the number of segments in the speech[3]. And $\sigma(.)$ here represents the similarity between two vectors (i.e., cosine distance) in a speaker space.

I assume that if there exist "easy speakers" whose voices are more discriminative than others, the difference between inter-speaker segment similarity and intra-speaker segment similarity should be large enough. Based on this assumption, the

[2]The i-vector of a cluster is usually the mean vector of all of the segment vectors included in that cluster.

[3]For segmentation, I used the same method as [88], which extracted segments every second using a 3 second analysis window.

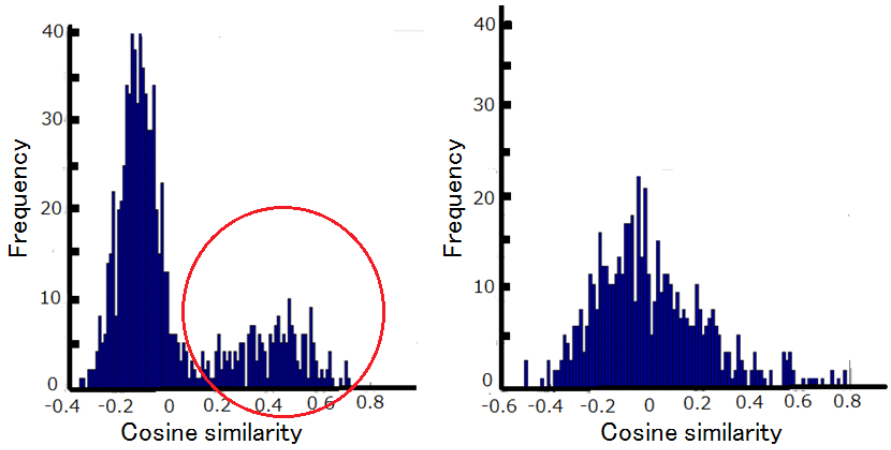

Figure 3.1: **The histograms of an ideal vector's(w_α) $\mathbf{D}_\alpha$ (left) and an imperfect vector's(w_β) $\mathbf{D}_\beta$ (right).** Segments in the red circle all belong to one speaker when clustering doing the trick.

distribution of $\mathbf{D}_\alpha$ should be distinctly bi-modal, with the within-speaker similarity having a mean/mode closer to 1 while the between-speaker scores have a mean/mode closer to 0. Moreover, as the extraction of i-vectors is based on the assumption that i-vectors from a specific speaker follow a Gaussian distribution[90]. We can then expect that if cluster α is an ideal cluster of an "easy speaker", the distribution of $\mathbf{D}_\alpha$ can be modelled by a two mixture Gaussian mixture models (GMMs), with one Gaussian composes within-cluster similarity scores and another Gaussian composes between-cluster similarity scores.

Therefore, I propose the use of two mixture Gaussian mixture models (GMMs) to model the distribution of $\mathbf{D}_\alpha$, and use the average log-likelihood of the GMMs to evaluate cluster C_α. The average log-likelihood evaluation can be written as follows:

$$\ln \hat{L} = \frac{1}{N} \sum_{i=1}^{N} \ln\{w_1 N(d_i; \mu_1, \sigma_1^2) + w_2 N(d_i; \mu_2, \sigma_2^2)\} \tag{3.4}$$

where d_i expresses elements belonging to $\mathbf{D}_\alpha$, and N is the number of segments. If cluster C_α is an ideal cluster, based on my assumption the distribution of $\mathbf{D}_\alpha$ tends to be bi-modal, which fits the two mixture GMM and leads to a larger in $\ln \hat{L}$. In contrast, if the distribution of $\mathbf{D}_\alpha$ is far from a bi-modal distribution, $\ln \hat{L}$ should be relatively smaller (see Fig. 3.1).

Finally, if the average likelihood is over the threshold, all the segments $\mathbf{s}_i$ which fit

$$\omega_\alpha N(d_i; \mu_\alpha, \sigma_\alpha^2) > \omega_{\bar{\alpha}} N(d_i; \mu_{\bar{\alpha}}, \sigma_{\bar{\alpha}}^2) \tag{3.5}$$

should be removed from the clustering process, considered as segments belonged to speaker A, and combined as one cluster. Here, μ_α, σ_α^2, and ω_α are the mean, variance, and weight of the Gaussian distribution with the larger mean, respectively; $\mu_{\bar{\alpha}}$, $\sigma_{\bar{\alpha}}^2$, and $\omega_{\bar{\alpha}}$ are the mean, variance, and weight of the Gaussian distribution with the smaller mean, respectively.

An alternative threshold has also been proposed, in which I ignore the influences of the variance and the weight. The threshold can be written as follows:

$$|\sigma(\mathbf{s}_i, w_\alpha) - \mu_\alpha| > |\sigma(\mathbf{s}_i, w_\alpha) - \mu_{\bar{\alpha}})|. \tag{3.6}$$

3.3 Modified bottom-up clustering

In this section, I describe the proposed modified bottom-up clustering method which uses the cluster evaluation method proposed in Section 3.2 to prevent over-merging (the merging of two different speakers' clusters). Bottom-up clustering is the most frequently used method of clustering when performing speaker diarization tasks [86]. The main approach of bottom-up clustering algorithm can be described as follows:

1. Divide all segments into K initial clusters;
2. Merge the cluster pair with the highest similarity score;
3. Re-build the cluster model, and perform re-segmentation based on the new model;
4. Repeat steps $2 - 3$ until a threshold for the highest similarity score for the merging of cluster pairs is met.

Conventional bottom-up clustering renews the cluster model with the entire input, calculates the similarity score of every cluster pair and merges the cluster pair with the highest similarity score in every iteration. Clustering finally stops when a threshold is met, and the remaining clusters, which are considered to be ideal

clusters for each speaker, are treated as the final output. Conventional bottom-up clustering methods label clusters (i.e., considers the clusters to be ideal clusters) only during its final stage, which means clusters containing segments from different speakers have the possibility of being over merged before the merging of clusters containing segments from the same speaker. As merging is not reversed during the process, this over-merging will result in a significant error. If we consider application, over-merging is obviously worse than under-merging (i.e., more than one cluster containing segments from the same speaker remain unmerged). Therefore, merging control is an important issue in speaker diarization, and some researchers even claim that it is necessary to set a relatively strict threshold to prevent over-merging [93].

The cluster evaluation method I proposed in Section 3.2 can detect speakers who are more easily identified at an earlier stage, which can prevent over-merging. In contrast to the conventional clustering methods, picking out some speakers at an early stage in the process allows us to set a more "limited" threshold which has potential ability to increase diarization accuracy. I propose a modified bottom-up clustering algorithm based on cluster evaluation, which can be described as follows:

1. Divide all segments into K initial clusters;

2. Merge the cluster pair with the highest similarity score;

3. Evaluate all the remaining clusters and remove any cluster(s) considered to be ideal from the clustering process;

4. Re-build the cluster model, and perform re-segmentation based on the new model

5. Repeat $2-4$ until a threshold for the highest similarity score for the merging of cluster pairs is met.

Because the final goal of speaker diarization is creating only one cluster for each speaker, clusters which are believed to be ideal no longer require further merging. Therefore, I proposed the use of cluster evaluation to detect clusters which are likely to be ideal and to remove any "ideal" clusters from further clustering.

3.4 Experiment

Eight meetings from AMI English meeting corpus [94] were used as development data, and eight other meetings from the same corpus were used as test data. The development data contained about 4.76 hours of meeting conversation, and the test data contained about 5.43 hours of meeting conversation. The AMI project is a multi-modal meeting analysis project which took place in Europe which created over one hundred hours of multi-modal meeting data including audio, video, annotation, etc., which is available for free download. The meetings were about production design, and $3-5$ participants discussed this theme during four meetings (one experimental set), each of which was about thirty minutes in length.

The meetings used in my experiment were basically randomly selected. However, since most of the meetings in the AMI corpus have four participants, I selected meetings with various numbers of participants in order to introduce variation in the number of speakers into the experiment. Thus, three meetings which contain three participants were selected directly[4]. Data recorded with a headset microphone was chosen. Development data contained eight meetings with four different participants in each meeting. Test data contained three meetings with three participants and five meetings with four participants. Speech characteristics (i.e., i-vectors) were extracted using ALIZE [95]. 60 dimensional linear frequency cepstral coefficients (LFCCs) were extracted every 10 ms using a 30 ms analysis window. Segments were extracted every second, using a 3 seconds analysis window. I-vectors were then computed for every segment using the LFCCs it contained. As my focus is on the proposed clustering evaluation method and modified clustering algorithm, in this experiment pre-processing, which includes voice activity detection (VAD) and overlapped speech detection, were performed manually.

3.4.1 Cluster evaluation

The experiment described in this section was performed to evaluate the cluster evaluation function proposed in Eq. (3.4).First, we need to develop a measurement to determine whether a cluster is an ideal cluster. The frame level (i.e., 10 ms)

[4]A meeting which contained five participants was not selected because one of the participants spoke too little (less than two minutes) during the meeting.

harmonic mean of the precision and recall (i.e., F-measure) of the main speaker (speaker who provided the most frames in a cluster) was used as the standard when assessing a cluster if it is ideal. Based on the conventional bottom-up clustering approach I introduced in Section 3.3, the development data were clustered from 16 random initial clusters into one cluster. Figure 3.2 shows the relationship between the ideal cluster measurement defined in this section and the cluster evaluation score proposed at Section 3.2, both of which were computed using all of the clusters created during conventional bottom- up clustering. There is clearly a positive relationship between the proposed ideal cluster measurement (i.e., the harmonic mean of the recall and precision of the cluster) and the proposed cluster evaluation function. Moreover, it seems that all of the data (all the clusters) can be divided into two groups at the position where the harmonic mean is about 0.8. As a result, I defined a cluster which has a harmonic mean of the recall and precision of its main speaker of over 0.8 as ideal cluster. Next, I investigated whether the proposed cluster evaluation function can detect these ideal clusters. In Figure 3.3, the recall, precision and F-value lines were plotted for this detection task. The highest F-value (about 0.9) appears when the average log-likelihood is about 0.32[5]. As a result, 0.3 was set to be the threshold for the proposed cluster evaluation method.

3.4.2 Modified bottom-up clustering

The initial number of clusters K was set to be 16 in the experiment. During initialization, the entire data was equally divided into K parts as K initial clusters [87][6]. Although i-vectors were used as the speech characteristic similarity measure in this experiment, I believe that the modified clustering method used here appropriates other similarity measures. K-means clustering [96] with the number of speakers as prior information, and conventional bottom-up clustering with stopping criterion trained using the development data were set to be the baselines. As the initial centroids affect the results of k-means clustering, all of the data was clustered using k-means twenty times with random initial centroids, and the best result was chosen

[5]I also tried setting the definition of "ideal cluster" as clusters with harmonic means larger than 0.7 and larger than 0.9. When clusters with harmonic means larger than 0.7 were defined as ideal, the highest F-value (0.95) occurs when the average log-likelihood is about 0.3; when clusters with harmonic mean larger than 0.9 were defined as ideal, the highest F-value (0.8) occurs when the average log-likelihood is about 0.4.

[6]I also tried setting K to 32, but the final results were almost the same.

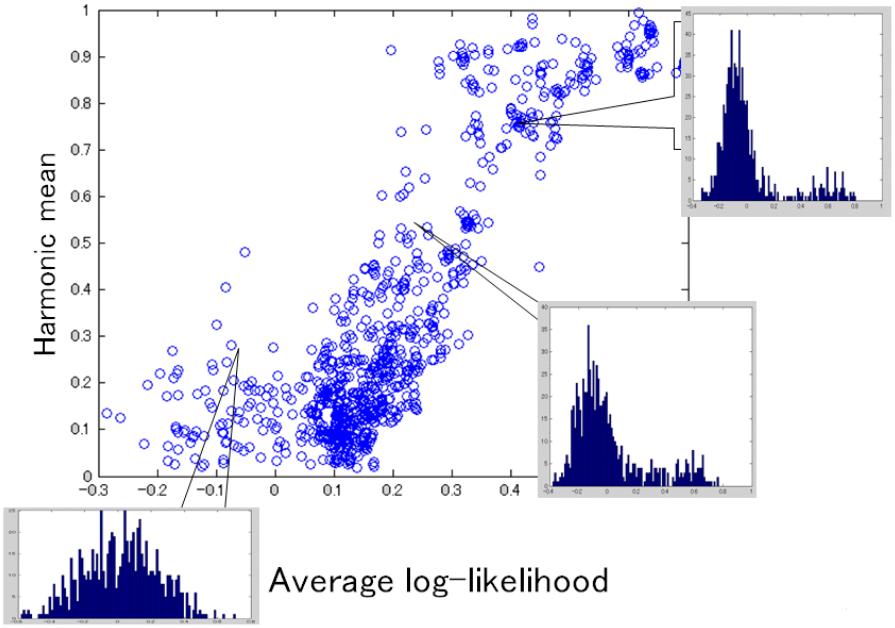

Figure 3.2: **Verification results of the proposed cluster evaluation method.** The horizontal axis is the average log-likelihood calculated using Eq. (3.4), and the vertical axis is the, harmonic mean of the recall and the precision of the dominant person's segments in the cluster.

as the baseline. The proposed method here means bottom-up clustering using the modified bottom-up clustering method proposed in Section 3.3 with the stopping criterion trained from the development data and cluster evaluation threshold defined in Section 3.4.1.

The method used to evaluate diarization accuracy when using the proposed method was based on [98], so-called the misclassification rate[7], cluster purity and Rand index were used.

Given a one-to-one speaker-to-cluster mapping, any frames (LFCCs) from speaker j that is not mapped is considered to be an error, the summation of which is denoted as e_j. Thus, misclassification rate defined as follows:

$$M = \frac{\sum_{j=1}^{N} e_j}{\sum_{i=1}^{C} \sum_{j=1}^{N} n_{ij}} \tag{3.7}$$

where n_{ij} represents frames from speaker j that are mapped to cluster i, C is the

[7]As the non-voice and overlapping speech sections were removed manually, the misclassification rate here is the same as the most popularly used diarization error rate (DER) [97].

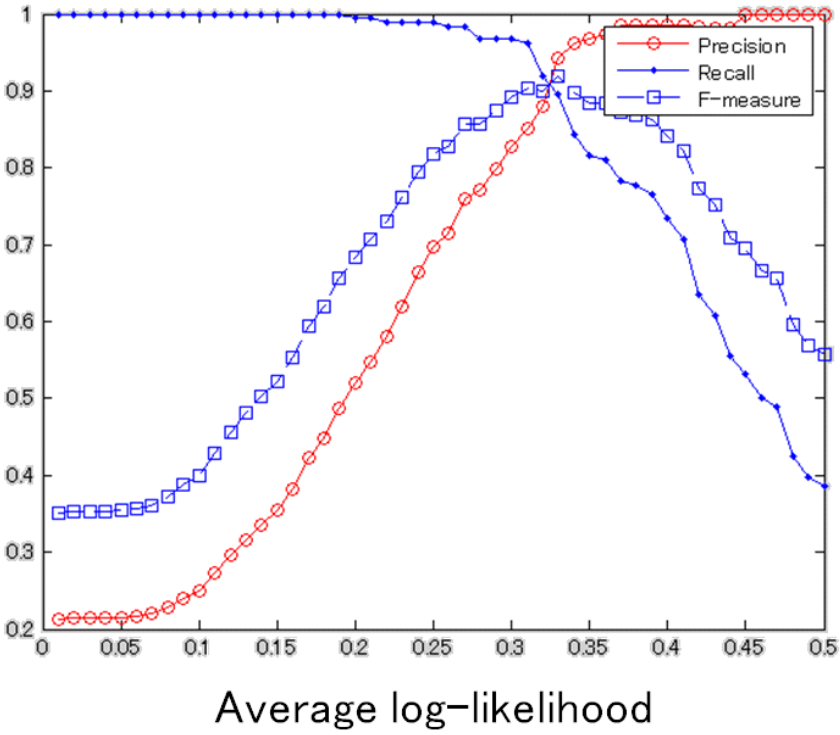

Figure 3.3: **F-measure of the proposed method when consider** 0.8 **as the threshold of the clusters' harmonic average (Fig.3.2)**

number of clusters, and N is the number of speakers. The final mapping was defined as the one that minimizes the misclassification rate.

Meanwhile, cluster purity was computed using:

$$P = \frac{\sum_{i=1}^{C} f_j}{\sum_{i=1}^{C} \sum_{j=1}^{N} n_{ij}} \tag{3.8}$$

where f_j is the number of frames which are the majority in cluster j.

Rand index has been proposed as a method to evaluate similarity between two clusters[99]. A Rand index can be computed as follows:

$$R = \frac{a+b}{\binom{n}{2}} \tag{3.9}$$

where a expresses the number of frame pairs which have been clustered into the same cluster during speaker diarization and in the ground truth, b expresses the number

Table 3.1: Diariazation result

Data	Misclassification rate (%)				Cluster purity (%)				Rand Index			
	CB	KM	PM1	PM2	CB	KM	PM1	PM2	CB	KM	PM1	PM2
EN2002c	9.11	8.65	9.11	9.11	91.00	91.46	91.00	91.00	0.89	0.89	0.89	0.89
EN2009b	21.44	7.34	15.90	15.94	88.81	92.85	84.75	84.25	0.84	0.90	0.82	0.82
IN1001	9.23	22.71	9.55	9.89	90.86	83.82	90.54	90.20	0.89	0.80	0.86	0.86
ES2003b	16.15	12.24	5.19	5.86	83.97	87.88	95.40	95.11	0.86	0.89	0.95	0.94
ES2007b	33.65	17.63	11.75	12.56	78.38	83.00	88.49	87.68	0.80	0.84	0.88	0.87
ES2008b	5.08	6.06	8.62	8.54	91.39	94.07	92.43	91.60	0.94	0.94	0.92	0.92
ES2014b	23.71	8.19	7.76	8.18	89.47	91.99	92.41	91.99	0.85	0.91	0.92	0.91
ES2016c	15.18	4.21	4.91	5.59	84.87	95.96	96.17	95.86	0.92	0.96	0.96	0.96
Ave.	16.69	10.81	9.10	9.46	87.04	90.14	91.34	90.96	0.87	0.89	0.90	0.89

of frame pairs which have been clustered into different clusters during speaker diarization and in the ground truth, and n is the number of frames.

Diarization results are shown in Table 3.1. Here CB and KM represent conventional bottom-up clustering and k-means clustering (baselines), respectively. PM1 and PM2 represent the proposed clustering methods using Eq. (3.5) and (3.6) as the cluster evaluation threshold, respectively. The result shows that the proposed methods achieve higher diarization accuracy than the baselines when using all three of the evaluation measures.

Furthermore, to evaluate whether the proposed method can help prevent over- merging, diarization results when using the proposed method and when using the conventional bottom-up clustering method with different stopping thresholds are shown in Figure 3.4. The results show that the proposed method is more robust for preventing over-merging than the conventional bottom-up clustering method, especially when the stopping threshold is small (i.e., hard to stop).

The original idea behind this study was that when doing speaker diarization tasks, the utterances of speakers who are easy to recognize should be removed first. But when all of the speakers are easy to recognize (e.g., ES2008b), or when cluster evaluation cannot extract "ideal clusters" (e.g., EN2002c), the proposed method delivers the same results as the conventional method. Moreover, similar to the conventional bottom-up method, if the cluster evaluation process mistakenly detects an "ideal cluster" (which did not happen in this experiment), a large error would occur as the clustering process cannot be reversed.

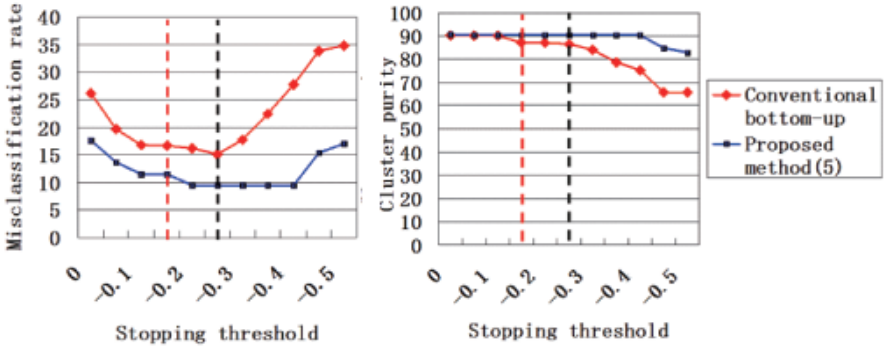

Figure 3.4: **The relationship between misclassification rate and the stopping threshold(left). The relationship between cluster purity and the stopping threshold(right).** The dashed line is the stopping threshold computed from the develop data.

3.5 Summary of this chapter

This chapter proposed a cluster evaluation method to judge whether a cluster is composed almost entirely of speech segments from only one speaker. Based on this proposed cluster evaluation method, a modified bottom-up clustering method was also proposed. The experimental results show that the proposed cluster evaluation method can detect "ideal clusters", and that the modified bottom-up clustering method achieved a higher level of diarization accuracy than the baseline methods.

The proposed method is based on statistical properties between within-cluster similarity score and between-cluster similarity score, hence it is less effective when applied to speakers who speak too little during the whole conversation (i.e., when there are not enough speech segments for that speaker). Moreover, although the experimental results showed that the proposed method can accurately detect "ideal clusters" (with an F-value of about 0.9), as I mentioned above, errors can result from the mistaken detection of "ideal clusters". These two problems will need to be solved in future research. Finally, as the number of speakers in the AMI meeting corpus has a relatively limited range, in order to test the generality of the proposed method, the experiment should be repeated under different diarization conditions.

Chapter 4

Impact of acoustic similarity on efficiency of linguistic information transmission via subtle prosodic cues

Keywords in this chapter:

- Subtle prosodic cues
- Prosody information transmission efficiency
- Voice morphing
- Eye-tracking
- Objective similarity measure

In this study I investigate the impact of acoustic differences on the efficiency of subtle prosodic information transmission. In the study, participants listened to lexically ambiguous sentences, which could be understood with prosodic cues, such as syllable length and pause length. Sentences were uttered in voices similar to the participant's own voice and in voices dissimilar to their own voice. The participants then identified which of four targets the speaker was referring to. Both the eye-movement and response time of the participants were recorded. Eye-tracking and response time results both showed that participants understood the lexically ambiguous sentences faster when listening to voices similar to their own. The results revealed that be similar in acoustic features, which do not contain linguistic information can influence the processing of linguistic information.

4.1 Introduction

Language comprehension involves a complex interaction between the transmitted message and the receiver's background knowledge and experiences [100]. As a result of this complexity, differences in representation styles can clearly influence the efficiency of our language comprehension process. For example, the inversion of subject and object in passive sentences makes these sentences more difficult for listeners to understand than sentences with the same meaning expressed using active form, for both positive and negative sentences [101]. Listeners also have difficulty interpreting "garden path" sentences, i.e., grammatically correct sentences which have meanings different from those that a listener would normally expect. For example, "The dog that I had really loved bones," and "I told her children are noisy." Such sentences are considered to be evidence of our sequential reading process (i.e., one word read at a time) [102].

Schema theory suggests that presenting messages in style that is familiar to the recipient improves linguistic comprehension efficiency, because when a receiver has relevant background knowledge, he or she can free up more working memory for analysis and interpretation of the linguistic information [68] [103]. Researchers have found evidence to support the theory that both lexical and prosodic familiarity increase the efficiency of our linguistic comprehension. Use of familiar topics has been found to help foreign language learners improve their performance on reading comprehension tasks, no matter which second language they are learning [104] or what their native language is [105]. Moreover, the facilitative effect of comprehension on language-related tasks is revealed in simple nativization drills, such as the changing of character and location names into native ones (e.g., when a Japanese English learner replaces "Barack Obama lives in Washington D.C." with "Shinzo Abe lives in Tokyo") [106]. Studies also show that familiarity with the speaker's prosodic speech characteristics, such as the speaker's accent, also have a positive influence on our listening comprehension, for both native and non-native listeners [107] [108].

If we want to apply this "speech characteristics familiarity effect" to provide high efficiency linguistic information transmitter (e.g., in call center), however, there are still questions need to be answered. One of the questions is how to measure the transmitter's speaker-specific speech characteristics familiarity to the listener. We can find that in most of the cases mentioned above, familiarity also involves self-

similarity (i.e., we are familiar with our own accent, capital, president, etc.). Thus, I propose self-similarity as a measure of speaker-specific speech characteristics familiarity, which is factor related with high efficiency communication. To examine this proposal, the present study conducted a behavioural experiments in which I tracked participants' eye movements while they listened to sentences and simultaneously watched related images on a computer screen and reacted. Images including lexically ambiguous items which can only be understood with subtle prosodic cues. Images were described at different levels of voice similarity to the participant, allowing me to investigate the impact of acoustic similarity on efficiency of subtle prosodic cues processing in real time.

This chapter is organized as follows. After a description of the experimental method, I describe my experimental procedure, report the experimental results and discuss their implications. And then end up with conclusions and a discussion of future research.

4.2 Method

This study employed lexically ambiguous material in the experiment to control the influence of lexical and prosodic features on comprehension. To vary similarity of the speakers' voices, morphing technology was used This allowed us to present information at different levels of self-similarity. This study also used objective similarity measures for further similarity analysis. To measure linguistic transmission efficiency, both response time during the target selection task and the proportion of the time participants were visually fixated on the appropriate target during the task were used.

4.2.1 Participants

Twenty-eight male, native Japanese-speaking college students were recruited as participants[1].

[1]I did not believe that gender would affect performance in this sort of comprehension experiment, and as a result there is an obvious imbalance in the genders of my participants. Future research should include more female participants, and should investigate the effect of a mixed-gender voice.

4.2.2 Materials

Spoken Japanese phrases with right-branching (RB) vs. left-branching (LB) ambiguities were employed as the experimental material. Fig. 4.1(a) shows an example[2]. In Japanese sentences such as "akai / hoshi no / nekutai" ("red (adjective phrase) / star (first noun phrase) / necktie (second noun phrase)") can be interpreted, as in English, as either "the red necktie with stars" or "the necktie with red stars". It is RB when the second phrase (the first noun phrase) should first be combined with the third phrase (the second noun phrase) (i.e. "the red necktie with stars"), and LB when the second phrase should first be combined with the first phrase (i.e. "the necktie with red stars"). These two phrases are identical in spelling and phonetic pronunciation, but can be distinguished by subtle prosodic cues [109]. No clear downstepping [3] "↘" from the first phrase to the second phrase, followed by downstepping "↘" from the second phrase to the third phrase suggests the right-branching meaning (the red necktie with stars), while clear downstepping "↘" from the first phrase to the second phrase, followed by moving up of pitch "↗" from the second phrase to the third phrase suggests the left-branching meaning (the necktie with red stars)[4]. A longer pause between the first and second phrases also indicates the RB meaning, while a longer pause between the second noun and its particle ("no"), inside the second phrase, indicates the LB meaning. A third prosodic cue is called "final segment duration", which is the duration of the final vowels in the different phrases. When the RB meaning is intended, there is longer final segment duration in the first phrase, while longer final segment duration in the second phrase implies the LB meaning (also see Fig. 4.2(a) and Fig. 4.2(b)).

[2]The other ambiguous material used in this study can be found in the appendix.

[3]A mechanism whereby the pitch register for marking accentual prominences, is lowered with each successive occurrence of a pitch accent within a phrase.

[4]Considered to be the main prosodic cue.

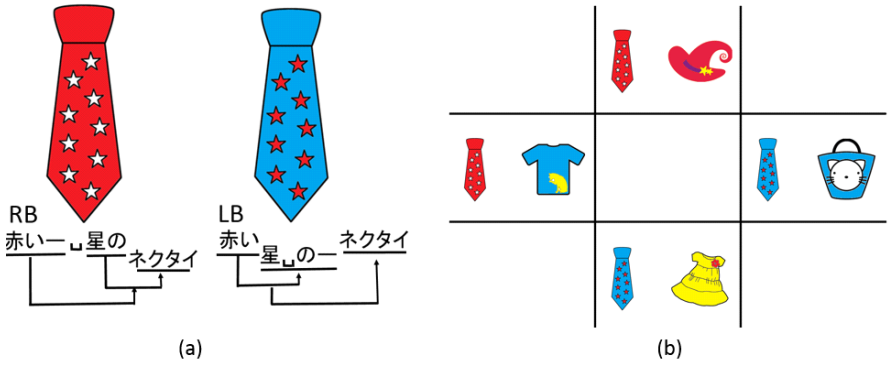

Figure 4.1: **Example of experimental items.** (a) Example of RB vs. LB ambiguity items used for recording; both of the pitcured items can be referred to as "akai hoshi no nekutai" in Japanese ("red star necktie" in English). RB prosodic cues: 1) No clear downstepping from the first phrase to the second phrase, followed by downstepping from the second phrase to the third phrase; 2) longer pause between the first and second phrases; 3) longer final segment duration in the first phrase. LB prosodic cues: 1) clearer downstepping from the first phrase to the second phrase, followed by moving up of pitch from the second phrase to the third phrase; 2) longer pause between the second noun and its particle ("no"), inside the second phrase; 3) longer final segment duration in the second phrase. In the figure, the lower height of a phrase means there is a clearer downstepping; a "␣" mark means there is a longer pause; a "-" mark means there is a longer final segment. And the pitch-height is indicated by a vertical placement of the text-characters. (b) Example of material used in each listening comprehension experiment trial.

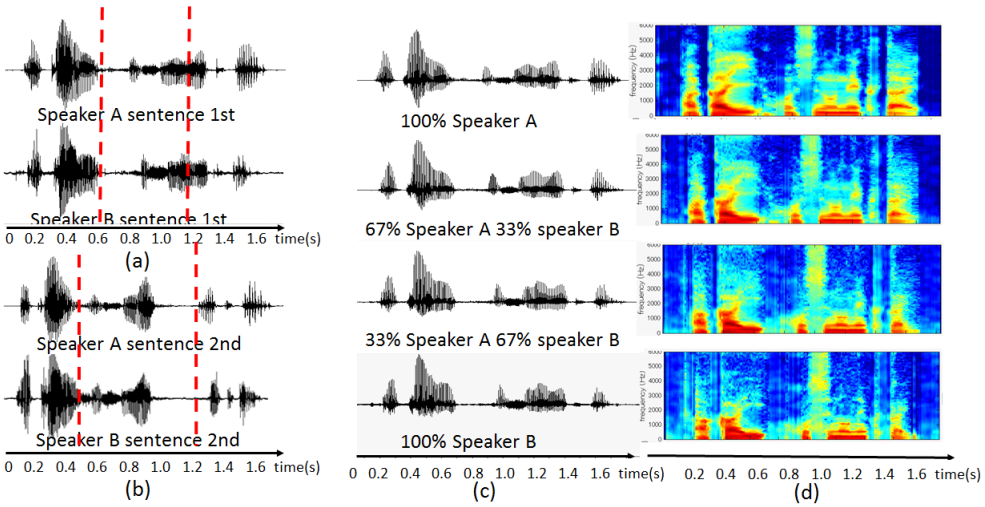

Figure 4.2: **Examples of different waveforms.** (a) Original waveforms of the phrase "the red necktie with stars" (RB) as read by different participants. (b) Original waveforms of "the necktie with red stars" (LB) as read by different participants. The dashed lines show the boundaries of each phrase in the upper sentence. (c) Synthesized waveforms when morphing the waveforms in (a) together under different morphing conditions. (d) Spectrogram information of waveforms shown in (c).

Sixty experimental pictures were created (see Fig. 4.3 for an example). Each picture consisted of four pairs of items that can be described using RB vs. LB phrase. In each picture, the two rectangles, which contained the correct first (left) item, no matter what the second item was, were defined as the "correct" areas, while the two rectangles, which contained the incorrect (ambiguous) first item were defined as the "incorrect" areas. Other parts of the screen, which had no items displayed were defined as "other" areas. In each trial the first item in the target pair was described ambiguously, while the second item in the target pair was described unambiguously. Note that until the description of the second item is provided, all of the rectangles containing the correct first item could be perceived by the participants as "correct" targets. In this experiment, I wanted to see whether there were differences in the proportion of visual fixation on "correct" pairs of items under different experimental conditions.

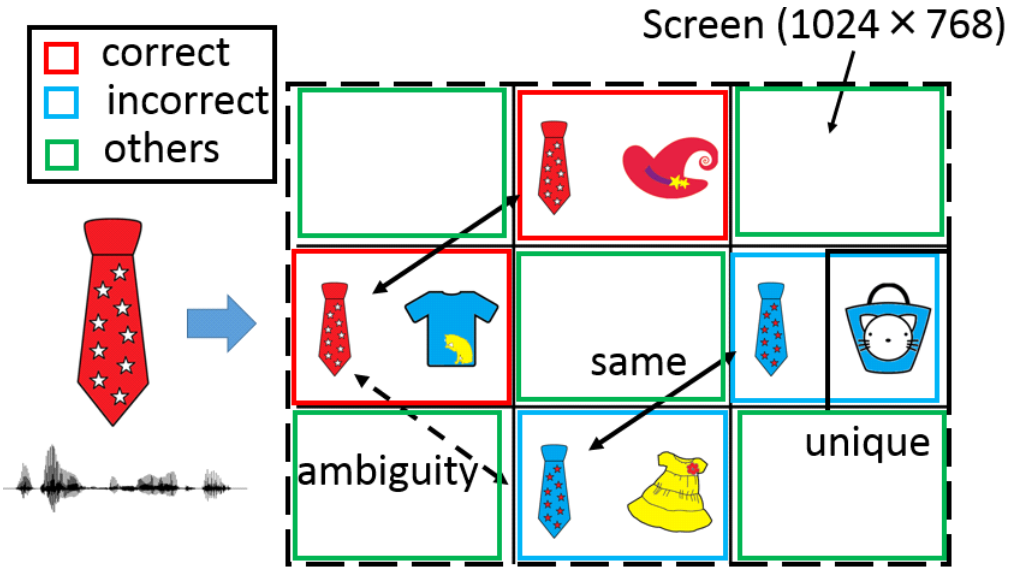

Figure 4.3: **Definition of visual fixation areas.** When the first item is described ambiguously but with prosodic cues as "red star necktie", the areas inside the red squares are defined as "correct" areas, while the areas inside the blue squares are defined as "incorrect" areas. The other areas of the screen are defined as "other" areas.

4.2.3 Voice morphing

Morphing techniques have been developed to change one stimulus object (e.g., an image) into another with a seamless transition. Since morphing techniques can enrich the level of stimulus without salient loss of naturalness, they have been used in many facial image-related experiments, such as those involving facial recognition [110] and attractiveness perception [111]. TANDEM-STRAIGHT [112] is a speech analysis, modification and re-synthesis framework, which can similarly deconstruct a speech signal based on the source-filter model. TANDEM-STRAIGHT extracts the F0 and aperiodicity of the input speech signal as the source parameters. The signal's spectrogram information was used together with its F0 to obtain the filter parameters. While morphing, the weighted average of all the parameters from the two source signals, which also included mapping information in the time and frequency domains, were used to re-synthesize the voice, based on the source-filter model[5] (see Fig. 4.4(a)). TANDEM-STRAIGHT can generate naturally sounding voices, allowing acoustic researchers to apply morphing techniques in their experiments in order to investigate the perception of paralinguistic and non-linguistic information in voices, such as the perception of gender [113] and speaker identification [114].

After the participants' voices were recorded reading the Japanese RB vs. LB ambiguous phrases, I randomly paired participants with a stranger[6] and used the TANDEM-STRAIGHT toolbox to morph their original voices into four transitional levels of similarity using manually anchored start and end points of each syllable. The starting point and ending point of each syllable were aligned manually (see Fig. 4.4(b), the white circles are the anchored points). The morphing conditions were; 100% speaker A's voice, 67% speaker A's voice mixed with 33% speaker B's voice, 33% speaker A's voice mixed with 67% speaker B's voice, and 100% speaker B's voice. As the synthesized voices still sound somewhat artificial, to compensate for this, voices were re-synthesized using TANDEM-STRAIGHT even for the 100% and 0% similarity conditions. Fig. 4.2(c) and (d) shows the morphed waveforms and spec-

[5]Although TANDEM-STRAIGHT allows users to modify the parameters independently (some of the parameters are fixed), however, in this study all of the parameters were modified together (i.e. replaced by a weighted average of the two source voices). This was because the main question I wanted to investigate was whether the similarity of interlocutor's voices influences information transmission.

[6]Before being paired-up with a partner, participants were shown a list of the names of all of the participants to make sure they did not know their partner.

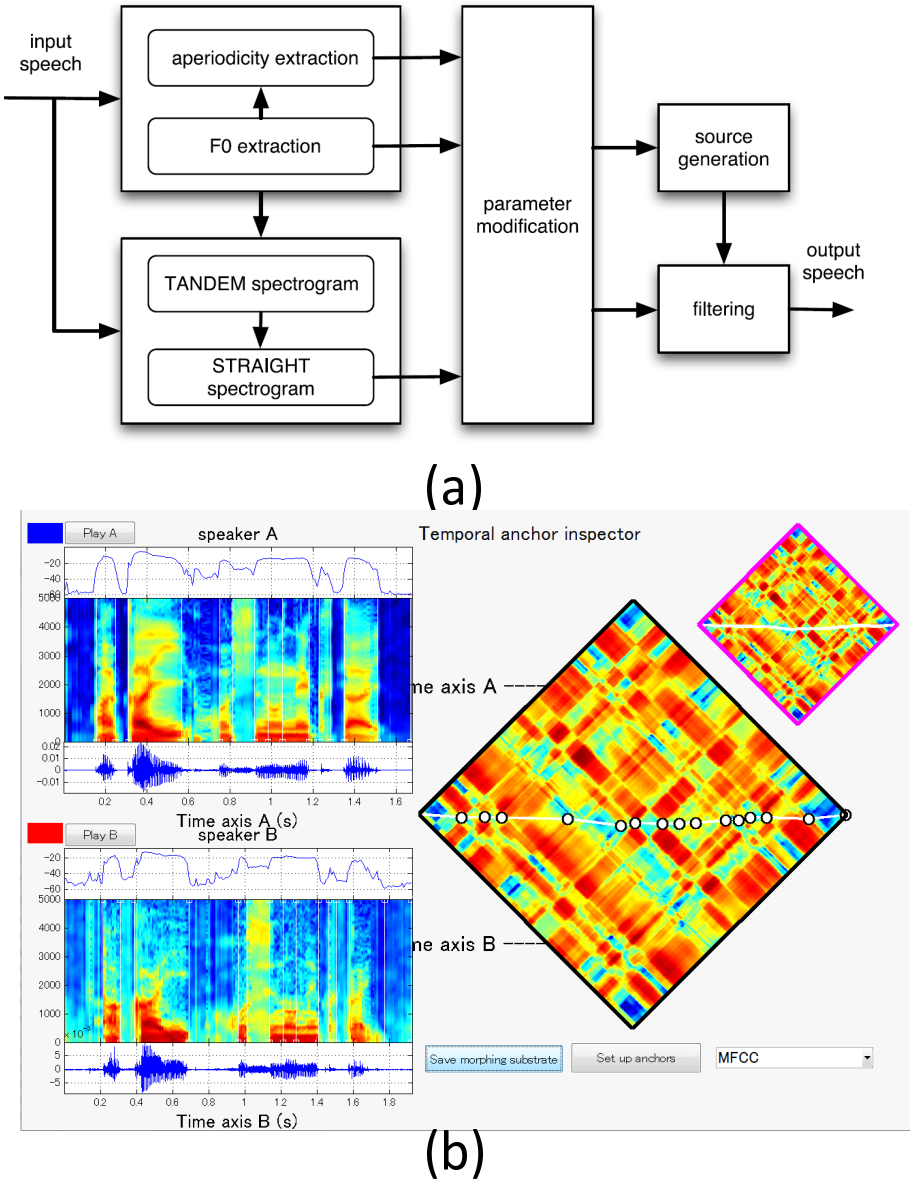

Figure 4.4: **TANDEM-STRAIGHT toolbox for voice morphing.** (a) Flow chart of TANDEM-STRAIGHT for voice synthesis. TANDEM-STRAIGHT extracts the F0 and aperiodicity of the input speech signal as the source parameters. The signal's spectrogram information was used together with its F0 to obtain the filter parameters. While morphing, the weighted average of all the parameters from the two source signals (also included other information such as mapping information in time and frequency domains) were used to re-synthesize the voice, based on the source-filter model. (b) Time anchor panel for voice morphing. The diagonally oriented square is the distance matrix of Signal A and Signal B. The white circles in the distance matrix are anchored points, which can be determined manually. White lines between anchored points show the aligned frames.

trum based on the waveforms shown in Fig. 4.2(a), respectively. And we can see that they are very similar to each other in timing and intensity.

4.2.4 Objective similarity measures

Although used morphing technology to artificially create voices with different levels of similarity, the original dissimilarity of the speaker's voices varied, i.e., for some participants, even in the 0% "own voice" condition (100% other person's voice), their partner's voice was still very similar to their own. Hence, I introduced objective similarity measures, which included spectrum, pitch contour and duration, to allow further analysis. The spectrum is assumed contains one's personal characteristics, which partially defines the acoustic features of an individual's speech. Meanwhile, prosodic cues, such as intonation and duration, are relevant to one's speaking style, which will also influence the acoustic features of one's speech. For convenience, all of these features are called "acoustic features" in this paper.

Spectrum similarity measure

The optimal cost of a dynamic time warping algorithm (DTW) is frequently used for measuring similarity between two spectral sequences. The DTW algorithm itself is used for measuring similarity between temporal sequences, based on a distance matrix and dynamic programming. In practice, DTW first evaluates the local alignment distance between each pair of elements in order to obtain a distance matrix. Then, a cost matrix is calculated from the distance matrix. The cost matrix is the same size as the distance matrix, with $C(1,1) = D(1,1)$ as its initial element, with

$$C(i,j) = D(i,j) + \min \left\{ \begin{array}{c} C(i\text{-}1,j) \\ C(i\text{-}1,j\text{-}1) \\ C(i,j\text{-}1) \end{array} \right\}, \tag{4.1}$$

as its other elements [7]). $D(i,j)$ is the entry of the local distance matrix and $C(i,j)$ is the entry of the cost matrix. Thus, the final entry in the cost matrix (e.g. $C(I,J)$) is the optimum global alignment cost. The optimum mapping path between the two input vectors can also be found by backtracking the optimum path of each node.

[7] There are numerous ways to calculate the cost matrix, and here only explain the method used in this paper (for more details see [115]).

In this paper, MFCC distance is used to compute the distance between each pair of spectra (one for partner A and one for partner B) for a given phrase (e.g. ''red star necktie") so that I can obtain a distance matrix.

After fixing the manually anchored points together, DTW is used to align the rest of the frames with each other. Spectrum information is extracted using TANDEM-STRAIGHT. MFCC distance, which is the logarithm of the Euclidean distance between two MFCC vectors normalized by the maximum value of the total Euclidean distance, is the default distance measurement for spectrum sequences employed by TANDEM-STRAIGHT (and the distance measurement recommended by its creators).

Pitch contour similarity measure

The weighted correlation proposed in [116] is used for measuring similarity between a pair of pitch contours. After aligning two speech segments using DTW (as explained in the previous subsection), their pitch contour similarity is then computed using the following formula:

$$r_{f_A,f_B} = \frac{\sum_{i=1}^{I} w(i)(f_A(i) - m_A)(f_B(i) - m_B)}{\sqrt{\sum_{i=1}^{n} w(i)(f_A(i) - m_A)^2 \sum_{i=1}^{n} w(i)(f_B(i) - m_B)^2}}, \tag{4.2}$$

where $f_A(i)$ and $f_B(i)$ represents the log F_0 [8] value of speaker A and B in the i-th aligned frame, respectively, m_A and m_B represent the mean log F_0 of speaker A and B in the current speech segment, respectively. I represents the number of frames in the aligned sequence, and $w(i)$ is the weighting factor, based on the frame signal power [9].

Duration similarity measure

The absolute mean difference between anchored intervals (in this case, representing syllable and pause duration) is used for measuring similarity between two sets of

[8] F_0 was tracked using TANDEM-STRAIGHT. Unvoiced intervals were interpolated based on a cost function aimed at minimizing discontinuities in the resulting trajectories and maximizing plausibility, based on the side information associated with F_0 candidates [117].

[9] In this paper, the signal power stands for the mean square of the input waveform.

anchored speech. After anchoring the start point and end point of each syllable manually, duration similarity is measured using the following formula

$$D_{S_A,S_B} = \frac{1}{N-1} \sum_{s=1}^{N-1} |S_A(s) - S_B(s)|, \tag{4.3}$$

where $S_A(s)$ and $S_B(s)$ are the s-th intervals of speaker A and B computed from the anchored points, respectively, and N is the number of anchored points.

4.2.5 Procedure

The experiment was divided into two phases. In the recording phase, participants were shown 13 pairs of pictures. The two pictures in each pair were different, but could be described using the same lexically ambiguous phrase, depending on whether the RB or LB reading was used. They were asked to describe each picture in Japanese twice, using their own natural speaking style, by reading the supplied ambiguous phrase. Example pictures and an example description are shown in Fig. 4.1(a). They were recorded in a sound-proof booth at 48,000 Hz with 20 bits sampling. Participants were then randomly paired with a stranger participant, and TANDEM-STRAIGHT was used to morph their voices with the voices of their partners.

In the second phase of the experiment, a listening comprehension experiment was performed about one week later. After completing two unambiguous warm-up trials, the only aim of which was to make sure that the participants understood what they should do during the experiment, participants listened to the previously recorded ambiguous phrases (in which their voices had been re-synthesized and morphed) while viewing images (1024×768 pixels) shown on a visual display (see Fig. 4.1(b)). Participants were asked to identify which target they heard described as quickly as possible by pressing one of four arrow keys on the keyboard. Note that participants listened to exactly the same phrases as their randomly paired stranger partner, the only difference being that the self-similarity conditions differed (i.e., one participant's voice was the "other's person's voice" for their partner, and vice versa).

During the experiment, the eye movements of the participants were tracked with a Tobii X2-30 eye tracker at a sampling frequency of 30Hz. The targets were pictures of pairs of items, all of which had been seen by the participants during the recording

phase of the experiment. The first (left) item in each pair was the subject of the ambiguous phrase, while the second (right) item was unique and was described without ambiguity. I included these "second items" because the prosodic differences between the descriptions of pairs of ambiguous options is very subtle. Based on previous research, even when listeners hear their own recorded voices, they can only achieve a comprehension accuracy of about 70%. By adding a unique "second item", I am able to better distinguish between confused responses (when the listener does not know which target is being described) and incorrect responses (when the listener presses the wrong key by mistake). Each set of four pairs of items included two pairs with correct first items and two pairs with incorrect first items, which could be easily mistaken for the correct items due to RB vs. LB ambiguity.

The listening comprehension experiment involved a total of sixty similar trials (i.e. 15 trials from the 26 in each morphing condition were randomly selected for each pair of participants). Fig. 4.5 shows an example of one trial. Participants were asked to select the correct pair of items based on the phrase they heard by using a keyboard. The phrases were a combination of the participants' morphed voices ("first item" and "second item") in the same morphing condition. As shown in Fig. 4.5, each trial was divided into four logical stages. The first stage was a five second preparation stage, in which the set of four picture pairs was shown without any sound. The second stage ran from the beginning of the description of the first item (the item on the left) to the end of the description of the first item. In the third stage, the participants heard the word "to" (pronounced like the word "toe" in English, which means "and" in Japanese) and then a 0.3 second pause. The fourth and final stage spanned the period from the beginning of the description of the second item until the participant's response via the keyboard [10].

[10]Participants can respond at any time during a trial, therefore the fourth stage is absent in some trials due to situations such as mistaken responses, etc.

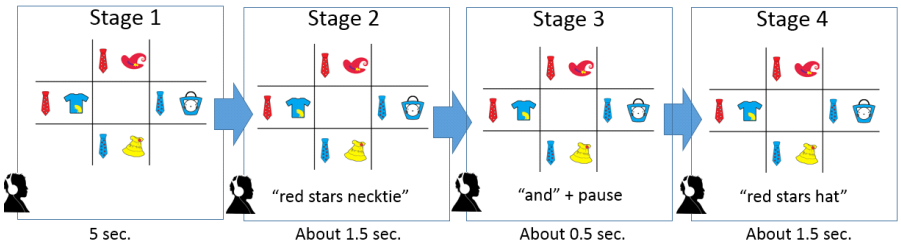

Figure 4.5: **Experimental procedure.** The experimental procedure was divided into four stages. In the first stage, only visual information is presented. During the second stage, information about the ambiguous item is presented. In the third stage, the word "to" (which corresponds to "and" in English) is heard, followed by a 0.3 second pause. In the fourth stage, information about the unique item is presented. The participant's comprehension of the ambiguous information is considered to occur during the second and third stage.

4.3 Results

In this paper, the results were analyzed using analysis of variance (ANOVA), which assumes that the ratio (i.e., F-value) of between-group variability to within-group variability follows an F-distribution. The probability (i.e., p-value) that the means of the experimental groups are all equal becomes smaller as the F-value increases. When the p-value is smaller than the alpha level (which was set to 0.05 for this study), the null hypothesis will be rejected (i.e., there is a significant difference between the means of the experimental performances of the groups being compared). Further, as four morphing levels were used in the experiment, Tukey's test was applied for pairwise comparisons when ANOVA shows that there is a significant difference in experimental performance.

The "stranger's voice" data was further divided into "strangers with voices similar to the listener's own voice" and "strangers with voices dissimilar to the listener's own voice" based on the objective similarity measures, which can be considered to be an extension of the original experiment. Thirty-third percentile and sixty-seventh percentile of all the data were set as thresholds for "similar stranger" and "dissimilar stranger", respectively. Participant pairs whose average objective similarity measure was higher or lower than these thresholds were considered to be a "similar stranger"

or "dissimilar stranger", respectively. Further ANOVA analysis was applied using the "similar stranger" and "dissimilar stranger" categories as an additional "between subjects" factor. Because I was afraid that similarity of pitch and duration of utterances within a participant pair could change (i.e., some utterances could sound similar while other utterances sounded dissimilar), for the purpose of analysis, both pitch and duration similarities were treated as both a "between subjects" factor and a "within subjects" factor (i.e., they were analyzed twice)[11]. Also note that there were only tiny differences in prosodic expression between paired participants. Fig. 4.6 shows the histograms of the similarity measures used in this study. The mean and variance of the mean differences in syllable and pause duration were $44.4ms$ and $378.04(ms)^2$, respectively. The mean and variance of the weighted correlation of pitch contours was 0.78[12] and 0.04, respectively.

[11]Participants/trials which did not meet both of the thresholds were ignored. For analysis of spectrum similarity, two participants were ignored. For pitch similarity, three participants were ignored. For duration similarity, four participants were ignored.

[12]A value that has been considered to indicate a high level perceptual prosodic similarity in previous researchs [116].

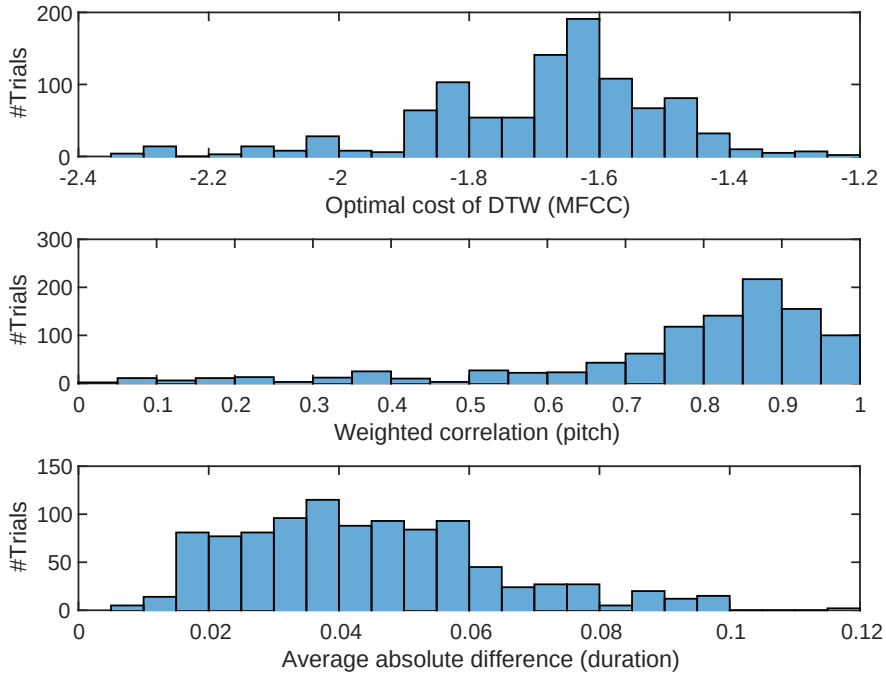

Figure 4.6: **Histograms of similarity measures used in this study.** The upper shows the histogram of MFCC similarity measure of all the trials. The middle shows the histogram of pitch similarity measure of all the trials.The lower shows the histogram of duration similarity measure of all the trials.

4.3.1 Pre-processing

Data collected from four of the participants was removed from analysis either because of experimental error (the participants misunderstood the task) or due to data recording error (50% of their eye movement data was lost). Thus, the study was conducted using data collected from twenty-four participants[13].

Although practice trials were conducted, there still appears to have been a strong practice effect in our results. Fig. 4.7 shows the average response time in chronological order. We can see a strong tendency toward decreasing response times as the experiment proceeds, especially at the beginning. Therefore, trials before the tenth trial were excluded from our results as training trials. This imbalance in the

[13]Trials in which participant gave an incorrect response or which had more than a 50% loss of eye movement data were also removed from analysis, which ignores 10% of the remaining data.

appearance of each morphing condition during pre-processing should be avoided in future research. Also, to control for individual differences in response times, the response times of each participant were normalized into z-scores for analysis as follows: $z = \frac{R-M_i}{\sigma_i}$ where R is the response time measured from the end of the speaker's production to the listener's keystroke response, M_i is the mean response time of participant i, and σ_i is the standard deviation of the response time of participant i.

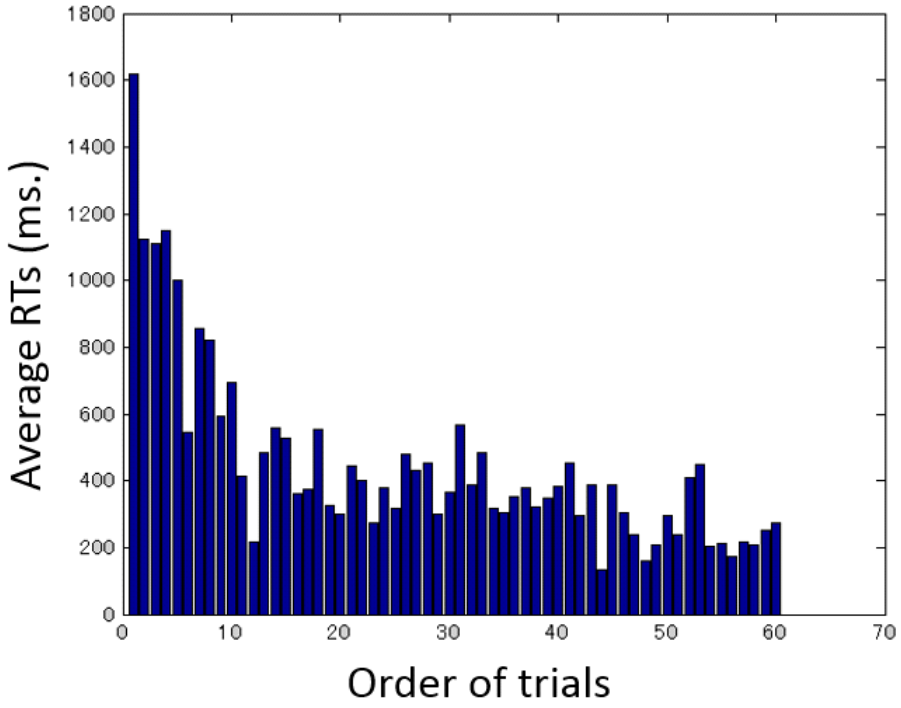

Figure 4.7: **Average response time for each trial.** The horizontal axis axis represents the order of the trials, while the vertical axis represents the average response time of the i-th trial from the end of the speaker's production to the listener's keystroke response.

4.3.2 Response time

Response time under different morphing conditions

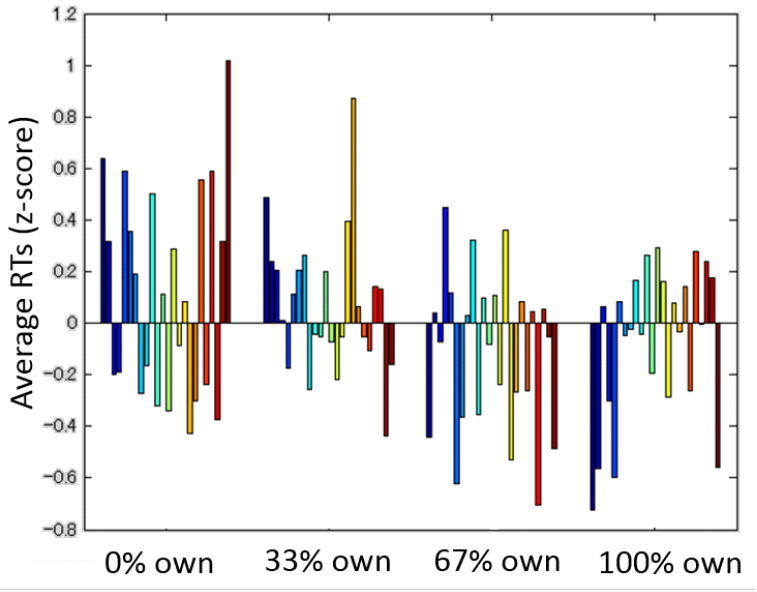

Figure 4.8: **Average response times of each participant under different voice morphing conditions.** Each coloured of bar shows one participant's average response time (z-score) under one morphing condition.

Fig. 4.8 shows the average normalized response times of each participant under different morphing conditions. Each color of bar shows one participant's average response time. We can see that when participants heard voices the same or similar to their own (100% own voice and 67% own voice) they responded faster than when they heard voices dissimilar to their own (33% own voice and 0% own voice). But little difference was observed between the 100% own voice and 67% own voice conditions, or between the 33% own voice and 0% own voice conditions. Statistical analysis also supported this observation. Tukey's test indicates that there are significant differences between the 100% own voice and both the 33% own voice and 0% own voice levels ($p < 0.01$), and also between the 67% own voice and both the 33% and 0% own voice levels ($p < 0.05$), but that there is no significant difference between the 100% own voice and 67% own voice levels, or between the 33% own voice and 0% own voice morphing conditions. It appeared that the participants could hardly distinguish the differences. I expected to find a linear relationship

between morphing level and perceived similarity, but this was not the case. Thus, the 100% own voice and the 67% own voice data were combined and considered both to represent the "own voice" condition, while the 33% own voice and 0% own voice data were similarly combined to represent the "stranger's voice" condition.

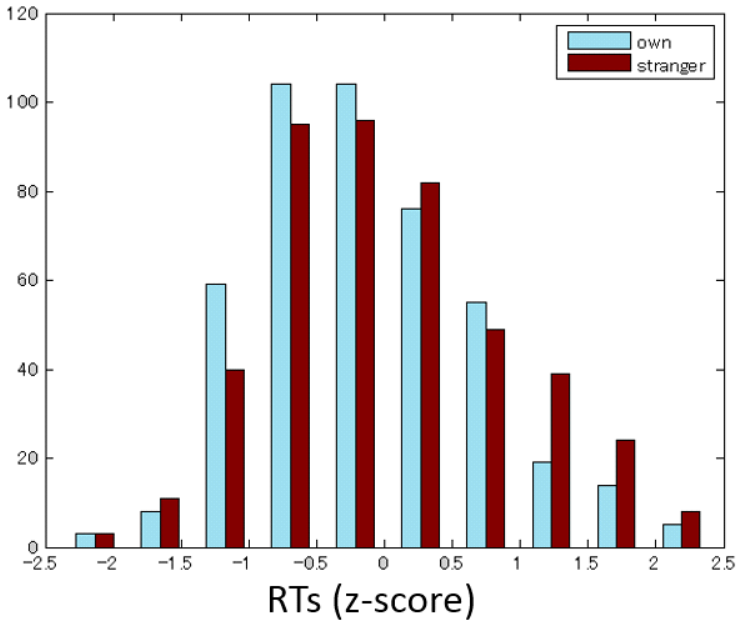

Figure 4.9: **Histogram of response times under different voice conditions.** Blue bars stand for the "stranger's voice" condition (67% stranger's voice and 100% stranger's voice), and red bars stand for the listener's own voice condition (67% own voice and 100% own voice). Horizontal axis represents the normalized (z-score) response time.

Fig 4.9, shows a histogram of normalized response times for the "own voice" and "stranger's voice" conditions using the combination of morphing data percentages described above. Similar to the results shown in Fig. 4.8, participants responded faster when prosodic information was presented in voices similar to their own. The morphing conditions were considered as within-subjects factor (designs), statistical analysis (ANOVA) shows a significant difference between these two groups of normalized response times ($F = 15.22$, $p < .001$). As both of the participants in each pair experienced exactly the same stimuli (saw the same pictures and heard the same voices), it should be able to exclude the possibility of irrelevant factors, such as the match-up between the images and spoken words, that may cause a difference in response times. Therefore, the significant difference in response times is probably

the result of the variation in the familiarity (similarity) of the voices presenting the information. For example, in Trial 1, Partner A heard his own voice describing the objects, while Partner B heard a stranger's voice (Partner A) describing the objects in his experiment.

Response time under different pairing conditions

There is still a significant difference between response times when using the duration similarity measure to divide "stranger" ($F = 7.754, p < 0.05$ as a between subjects factor, $F = 3.37, p < 0.05$ as a within subjects factor). However, there is no significant difference in response time between trials divided by spectrum similarity measure ($F = 2.10, p = 0.16$) or pitch similarity measure ($F = 1.55, p = 0.23$ as a within subjects factor, $F = 1.1, p = 0.34$ as a between subjects factor). One possible explanation is that differences in prosodic information comprehension are difficult to catch using response time as an indicator, and the difference in duration itself causes different response times (e.g. one's response would probably be slower when the stimulus lasts longer).

4.3.3 Degree of visual fixation

Visual fixation under different voice morphing conditions

Fig. 4.10 shows the proportion of listener eye fixation on the "correct" areas of the screen under different morphing conditions. We can see that although there is little difference during the second stage of the trial, participants were more likely to focus on the "correct" target during the third stage when the voice they were listening to was more similar to their own voice. The second stage includes the period from the beginning to the end of the description of the first item, and the third stage is listening to the word "and " followed by a short pause. These results support our hypothesis that listeners can more easily catch the subtle prosodic cues which help them to resolve lexical ambiguity when they are listening to voices similar their own. There was no statistical difference between eye fixation on the "correct" and "incorrect" areas of the diagrams by the participants to confirm this, however.

f

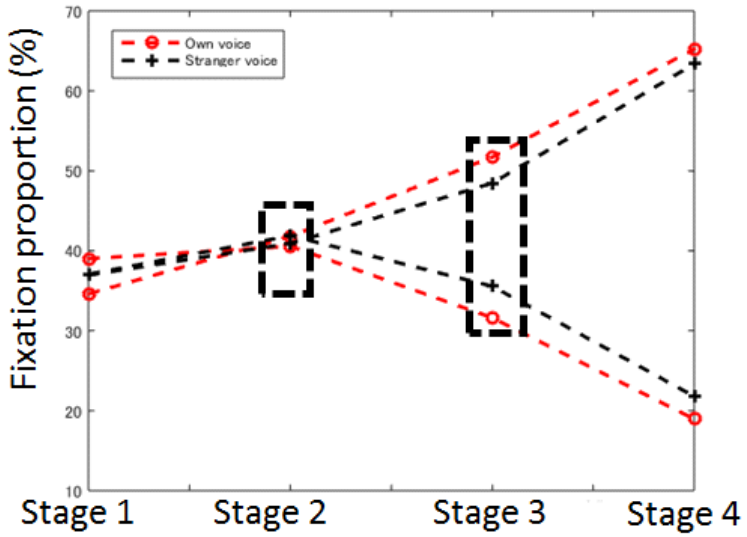

Figure 4.10: **Proportion of visual fixation on correct/incorrect areas under different morphing conditions during each stage of experimental trials.** The upper red line shows the proportion of visual fixation on the area of the correct first item under the "own voice" condition (67% own voice and 100% own voice). The lower red line shows the proportion of visual fixation on areas of incorrect first items under the "own voice" condition. The upper black line shows the proportion of visual fixation on the area of the correct first item under the "stranger's voice" condition (67% stranger's voice and 100% stranger's voice). The lower black line shows the proportion of visual fixation on incorrect areas under the "stranger's voice" condition.

Visual fixation under different pairing conditions

Just as in the previous section regarding response time under different pairing conditions, the "stranger's voice" condition was further divided into other voices similar to the listener's voice and other voices dissimilar to the listener's voice, and investigated differences in the visual fixation of the participants. Fig. 4.11 shows the proportion of visual fixation on the "correct" areas under different spectrum similarity levels. Fig. 4.12 shows the proportion of visual fixation on the "correct" areas under different pitch contour similarity levels (considered as within subjects factor). Fig. 4.13 shows the proportion of visual fixation on the "correct" areas under different syllable/pause duration similarity levels (considered as within subjects factor)[14].

[14]Here I only show the proportion of visual fixation on the "correct" areas for simplicity.

From these three figures We can see that when the listener hears another person's voice, which is similar to their own, their visual fixation during the third stage of the trials is the same as when they are listening to their own voice, especially when the trials are analyzed using spectrum and pitch contour similarity measurements. On the other hand, when listeners heard the voices of others, which differed from their own voices, We can see that their visual activity was more chaotic when selecting a fixation target. Statistical analysis shows a significant difference in the proportion of visual fixation on "correct" areas of the target when the "stranger's voices" were divided by spectrum similarity measure ($F = 4.64, p < 0.05$) and pitch contour similarity measure ($F = 8.32, p < 0.01$ as a between subjects factor, $F = 3.51, p < 0.05$ as a within subjects factor). Also note that in Fig. 4.13, while visual fixation on the "correct" areas under the "own voice" conditions and "similar stranger's voice" conditions are still similar, in contrast in Fig. 4.11 and Fig. 4.12, we can see that the proportion of visual fixation on "correct" areas is lower during the third stage under the "dissimilar stranger's voice" condition ($F = 0.36, p = 0.55$ as a between subjects factor, $F = 1.34, p = 0.27$ as a within subjects factor). This result may be because the duration cues used by different participants were perceptually more similar than the other two cues (i.e., changes in pitch and spectrum).

In summary, since the audio stimuli used in these experiments were verbally identical, the results of our experiment indicate that similarity in subtle prosodic cues does indeed positively influence the efficiency of prosodic information transmission. Additionally, there are significant differences in response times at different morphing levels and under different duration-based pairing conditions, but no significant difference in response times between MFCC-based pairing conditions or pitch contour based pairing conditions. In contrast, the visual fixation results show no significant differences at different morphing levels or different duration-based pairing conditions, but show significant differences between different MFCC-based pairing conditions and pitch contour based pairing conditions. I cannot explain this contrastive result, except to suggest that perhaps this experiment revealed a "boundary" of human speech perception ability. Investigation of a possible boundary of this type would be an interesting topic of future research. Also note that the utterances of some pairs of participants may have sounded more artificial than others, and that even within the same pair of participants some sentences sounded more artificial than others since nasal sounds usually sound slightly more artificial than plosive

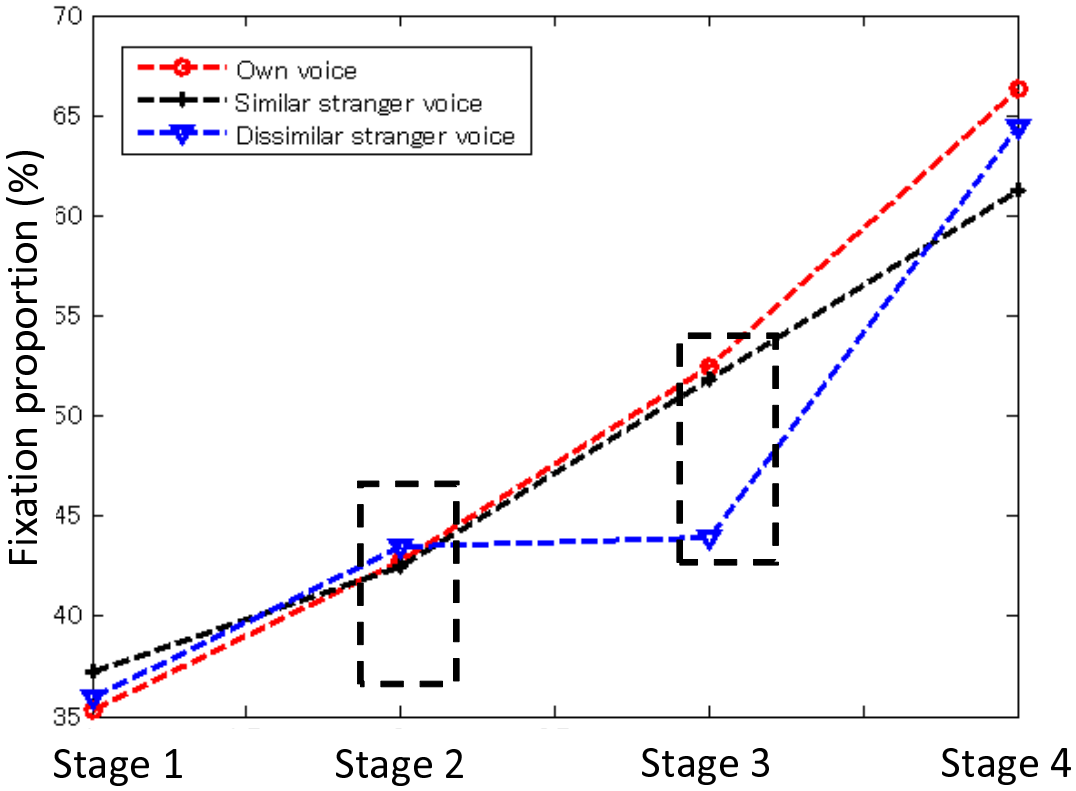

Figure 4.11: **Proportion of visual fixation on correct areas under different similarity conditions (DTW cost) during different trial stages.** Red shows the proportion of visual fixation on areas with the correct first item under the "own voice" condition (same as in Fig. 4.10). Black shows the proportion of visual fixation on areas with the correct first item under the "similar stranger's voice" condition. Blue shows the proportion of visual fixation on areas with the correct first item under the "dissimilar stranger's voice" condition.

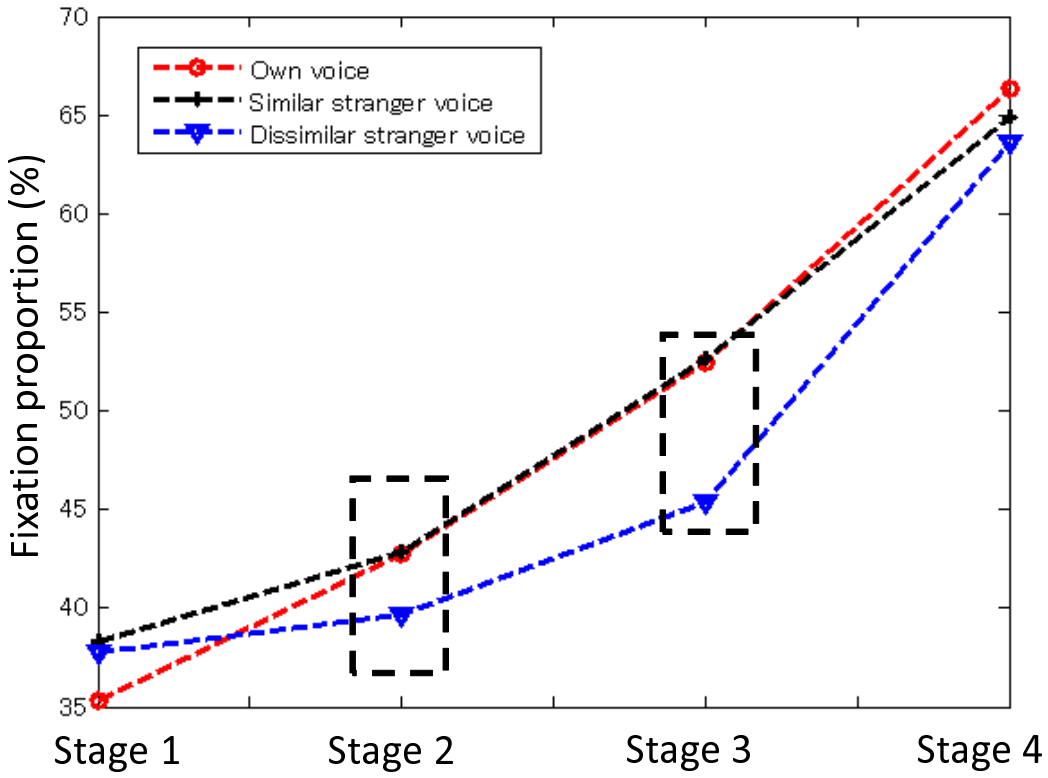

Figure 4.12: **Proportion of visual fixation on correct areas under different similarity conditions (pitch contour) during various trial stages.** Red shows the proportion of visual fixation on areas with the correct first item under the "own voice" condition (same as in Fig. 4.10). Black shows the proportion of visual fixation on areas with the correct first item under the "similar stranger's voice" condition. Blue shows the proportion of visual fixation on areas with the correct first item under the "dissimilar stranger's voice" condition.

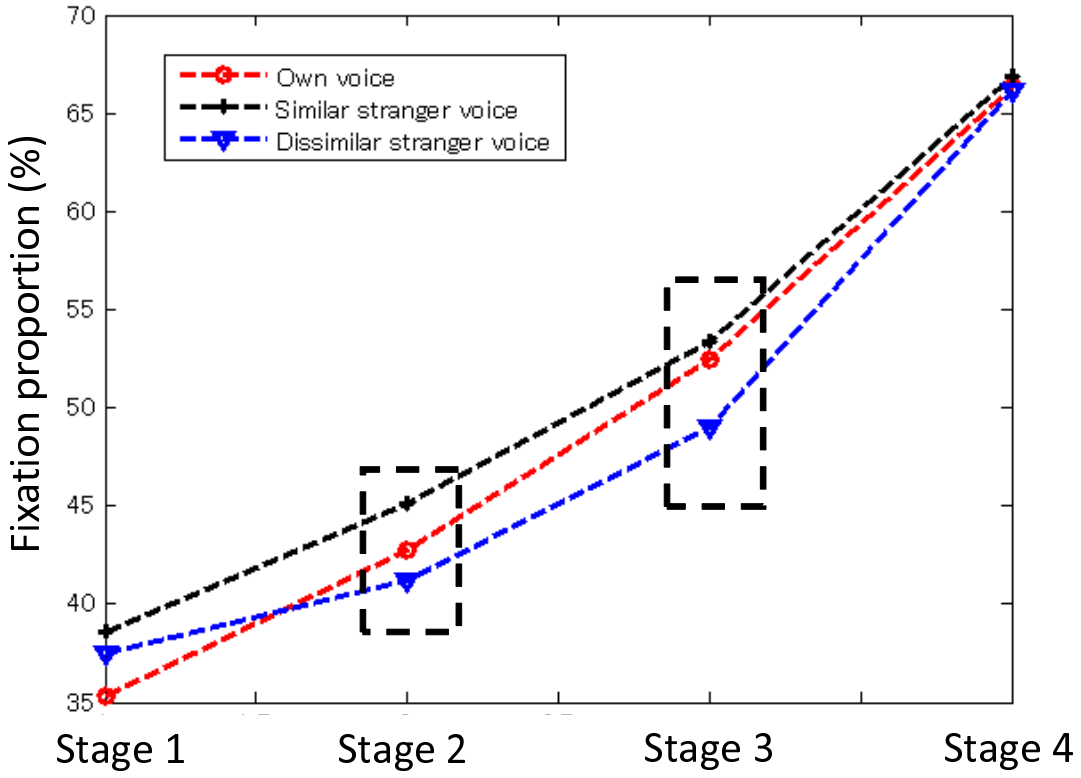

Figure 4.13: **Proportion of visual fixation on correct areas under different similarity conditions (duration) during various trial stages.** Red shows the proportion of visual fixation on areas with the correct first item under the "own voice" condition (same as in Fig. 4.10). Black shows the proportion of visual fixation on areas with the correct first item under the "similar stranger's voice" condition. Blue shows the proportion of visual fixation on areas with the correct first item under the "dissimilar stranger's voice" condition.

sounds. This research does not investigate the influence of the naturalness of the synthesized voices, which should also be examined in future research.

4.4 Summary of this chapter

Experiments to investigate the effect of subtle prosodic similarity on the efficiency of prosodic information transmission were designed and conducted in this study. Sentences with RB vs. LB ambiguity were used as the experimental material, and voice morphing technology was used to control voice similarity levels during the experiments. Objective similarity measurements were also used for analysis. Participants' response times and visual fixation behaviour were recorded. Analysis of the response time data showed that participants identified ambiguous target images more quickly when they heard voices similar to their own. Analysis of the visual fixation data also showed that participants understood more of the prosodically conveyed information when the target images were described in voices similar to their own. The results support the hypotheses that similarity in the speech characteristics of the information sender and information receiver result in higher information transmission efficiency, and that subtle acoustic cues influence efficiency of information transmission.

These findings were consistent with one another and imply that acoustic feature similarity is relevant to prosodic information transmission efficiency. In comparison to previous researches, the subjects of this study were all male undergraduate students who were native speakers of standard Japanese. The results suggest that human processing of speech information is so sensitive that even subtle prosodic cues influence our information transmission efficiency and language processing ability. But it should also be noted that only half of our experimental results were statistically significant, thus additional experiments which can verify the findings and investigate the "boundary" of human speech perception ability are needed. Finally, as spectrum similarity (MFCC distance) is considered to contain information on the condition of the vocal tract, the results suggest that physiological similarity is likely to be an additional dimension which needs to be considered when discussing speech communication and information transmission between speakers.

Regarding future works, the current experiment is unbalance in participants' gen-

der and the appearance of different morphing conditions, a stricter experiment with female participants ought to be done in the future. Also, as mentioned above, synthesized voices still sound somewhat artificial. Therefore, further investigation of the naturalness of morphed stimuli and their impact on information transmission is a potential area of research. Furthermore, instead of using morphed stimuli, information transmission efficiency when using "similar" or "dissimilar" participants' voices, as determined through the use of an objective similarity measure, should also be investigated. The combination of these two research projects might help us to verify that the slower listener reactions are not merely due to lower-quality stimuli or the amount of morphing, or due to the possibility that participants can identify their own voices and therefore exert extra effort.

Chapter 5

Correlation between Similarity in Speech Characteristics and Information Transmission Quality in Spoken Dialogue

Keywords in this chapter:

- Speaker similarity
- Speech characteristics
- Information transmission quality
- Map task

This study proposed speech characteristics similarity as a predictor of information transmission quality in spoken dialogue, and validates the proposal using a map task dialogue corpus. Typical speech characteristics similarity measures proposed for automatic speaker recognition based on their segmental features, prosodic features and idiolect features had been chosen. Likelihood ratio test of generalized mixed linear model was used to evaluate whether the selected similarity measures are effective predictors of the map task success. The results revealed an facilitative effect, with more successful in dialogues occurred between similar interlocutors than in dialogues occurred between dissimilar interlocutors.

5.1 Introduction

Speech signal we heard not only contains linguistic content, such as syllables, words, and phrases of the current utterance, but also ample amount of speaker-specific speech characteristics, such as fundamental frequency, formant spacing, and overall speaking rate. It means that when processing speech signal, listener should somehow process linguistic information as well as speaker-specific information. Previous researches suggest that these two process interact with each other, which means that speaker variation affects linguistic process[75][76][118][119], and vice versa[120][121][122]. For example in [123], compared to participants who were asked to do nothing during training phase, participants who were asked to finish a word identification task and participants who were asked to finish a speaker identification task during training phase both showed better (and similar) performance in a later phrases transcription test. As a result, it is natural for us to consider refining speaker-specific speech characteristics as an approach towards better linguistic information transmission quality. Increasing listener familiarity with speaker-specific speech characteristics has been found to be a potential direction of such a refining. It is found that not only be familiar with linguistic content related speech characteristics, such as accent and intonation, can facilitate speech perception process [107][124][125][126], but also speech characteristics that considered to be less relevant to linguistic content, such as the overall fundamental frequency. have facilitatory effects on the speech perception process[127][128]. For example in [129], researchers found that participants have higher accuracy in a word identification task, which was considered as a measure of their implicit memory, when the intonation contour, emotional prosody, or fundamental frequency of the voice in test phase are the same as what they had heard during training phase.

If we want to apply this "speech characteristics familiarity effect" to provide high quality linguistic information transmitter (e.g., in call center), however, there are still questions need to be answered. One of the questions is how to measure the transmitter's speaker-specific speech characteristics familiarity to the listener. In my last study, I proposed similarity in speech characteristics between listener and speaker as a measure of familiarity. Although similarity it is not exactly familiarity, most of us will accept that we are familiar with our own accent, intonation, vocabulary, culture, and so forth. This proposal has been validated using an am-

biguous sentences interpretation experiment in voice morphing similarity measure and objective similarity measure [130]. Results showed that when participants hear voices similar to their own at the segmental and prosodic (suprasegmental) levels, they could understand subtle prosodic cues better and faster, reducing ambiguity. Another question that need to be answered is whether it is still operative in the processing of spontaneous dialogues. As most of the previous studies relying on carefully controlled behavioural experiments, whether familiarity with speaker-specific speech characteristics affects linguistic information transmission quality is unclear. The present study investigated this issue using a map task dialogue corpus. Several typical speech characteristics similarity measures had been chosen to evaluate the impact of segmental level, prosodic level and idiolect level speaker/listener similarity on information transmission quality.

This chapter is organized as follows. After a brief introduction to the map task corpus and its information transmission quality measure used in this study in Section 5.2, I describe the selected methods of measuring similarity in speech characteristics in Section 5.3. In Section 5.4, I report the experimental results and discuss their implications. I end the chapter with conclusions and a discussion of my future research.

5.2 Human Communication Research Centre (HCRC) map task corpus [131]

The HCRC Map Task Corpus is a set of 128 direction sharing dialogues which have been recorded, transcribed, and annotated. The dialogues have been released and have been used by researchers investigating a wide range of behaviors [131]. There are 64 speakers featured in the corpus, all of whom were born in Scotland, who each take part in four conversations. The main task for the subjects is route guidance as shown in Fig. 5.1. Pairs of two participants take turns playing the roles of a giver and a follower of directions. During the dialogue, the direction giver describes a route that appears on his or her own map to the direction follower, using speech communication only, and the direction follower then tries to reproduce the same route on his or her own map. After the dialogue is finished, both the direction giver's and direction follower's A3 sized maps are covered with a grid of 1 cm squares, and

the deviation between their routes, measured in squares, is then calculated. This path deviation value is then used to measure the linguistic information transmission quality in this study. Note that, half of the dialogues occurred between familiar (i.e., acquainted) interlocutors, while the other half occurred between unfamiliar interlocutors. Meanwhile, half of the participants who took part in the task were able to make eye-contact with their partner, while the other half had no eye-contact. These experimental conditions, together with the participants' gender information will be used to set up a baseline prediction model in the experiment section.

5.3 Measures of similarity in speaker-specific speech characteristics

Typical speaker-specific speech characteristics similarity measures proposed for automatic speaker recognition have been chosen to examine whether be similar in segmental level , prosodic level , and idiolect level speaker-specific speech characteristics improve the linguistic information transmission quality between listener and speaker. As most of them are well known and operative in general automatic speaker recognition task, they are considered to have enough discriminative capacity to measure the similarity between two pieces of speech signals.

5.3.1 Segmental similarity measures

Here, segmental features refer to the features that calculated from the smallest discrete linguistic unit that can be identified. In practice, segmental features particularly refer to information that extracted from a short-term analysis window (usually $20ms$ $30ms$) which is considered to contain phonemic information. I-vector of MFCCs and some segmental features' overall statistical values were chosen to be the segmental similarity measures of this study.

MFCCs (mel frequency cepstral coefficients), which are believed to contain formant information about the human voice in a small piece of speech (usually $20ms$ $30ms$), are the most popular segmental features used in both speaker identification and speech recognition research. I-vector, which has state-of-art performance in speaker related recognition task (e.g., speaker recognition, speaker identification, speaker

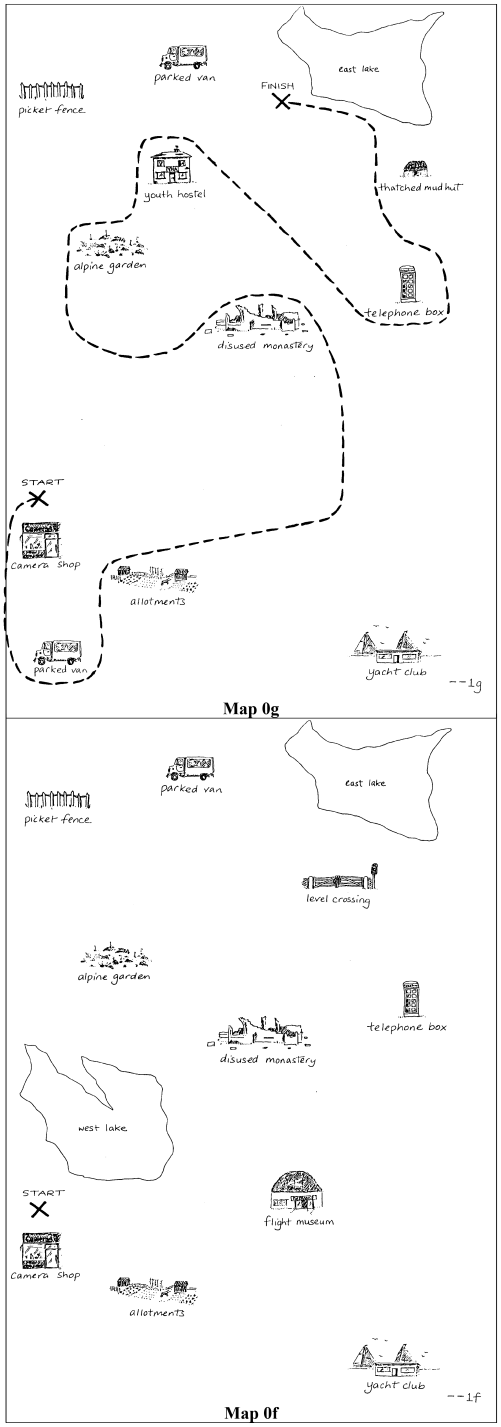

Figure 5.1: **Example of HCRC map task maps.** Note that to vary task difficulty, the direction giver's and the direction follower's maps were somewhat different in appearance and labelling.

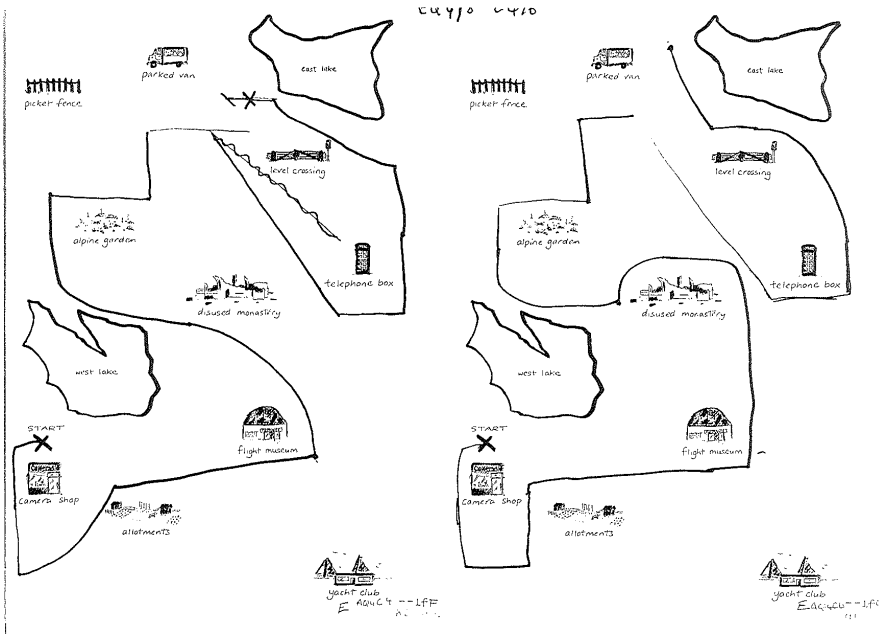

Figure 5.2: **Examples of routes drawn by direction follower.** The path deviations calculated for the completed route for q4ec4 (left) and q4ec8 (right) were 73 and 30, respectively.

verification, etc.) and has already become source features of some speaker recognition challenges[132], was used to calculate the similarity between giver's and follower's MFCCs,

Because the acoustic stream contains not only speaker identity information but also other verbal information, When doing speaker recognition, we want to exclude any information which may influence our evaluation of the similarity in speaker-specific speech characteristics. An i-vector is created to perform this task using a specified joint factor analysis. Based on the procedure used in [90], a speaker- and channel- (or session-) dependent supervector M, which is usually composed of the mean vectors of the Gaussians in the GMM of the current data set, can be expressed as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{5.1}$$

where $\mathbf{m}$ is the speaker- and channel-independent supervector, commonly taken from a universal background GMM representing speaker- and channel-independent acoustic features. $\mathbf{T}$ is a rectangular matrix using virtual speakers as its dimensions (i.e., it is the speaker space), and $\mathbf{w}$ is a vector called the total factor vector or

I-vector.

Cosine similarity has been applied successfully in the speaker space to measure the similarity of two I-vectors. Therefore, the similarity of two data sets, X and Y, represented by two vectors in the Total Variability space (the speaker space) $\mathbf{w}_X$ and $\mathbf{w}_Y$ respectively, can be measured using the following equation:

$$\sigma(\mathbf{w}_X, \mathbf{w}_Y) = \frac{(\mathbf{w}_X)^t(\mathbf{w}_Y)}{||\mathbf{w}_X|| \cdot ||\mathbf{w}_Y||}, \tag{5.2}$$

where, X and Y represent, respectively, all the giver speech segments of the current direction giver and all the giver speech segments of the current direction follower when he or she played the role of direction giver.

5.3.2 Prosodic similarity measures

Prosodic (or suprasegmental) features refer to features that extend over syllables and longer regions, they are usually considered as complementary information to automatic speaker recognition systems based on segmental features.

In [133], researchers used the slope of both pitch and intensity contours, in conjunction with segment duration to encode the prosodic dynamic features (or "pitch accent" in the words of the authors) of speech segment. After a speech segment is divided into syllable-like units by an automatic syllable detector, the slope of both the pitch and intensity of the units were calculated. Positive slopes were coded as "+" and negative slopes were coded as "−". In addition, segment durations were coded as either **S**, **L**, or **M**, with **S** representing the shortest 33% of segment durations, **L** representing the longest 33% of segment durations, and **M** representing all segment durations in between. This means that for each syllable-like speech unit, three symbols will be used to encode the slope of the pitch contour, the slope of the intensity contour and the duration of the speech segment (e.g., ++S or $+-M$). A bigram model is then used to model all of the coded segments spoken by a particular speaker. To measure the similarity of two bigram models constructed from the speech of the direction giver and direction follower in a dialogue, respectively, This study used a conventional log likelihood ratio test, which was also used in the original paper [133] and [51], to measure the similarity between two bigram models.

The similarity score can be expressed as:

$$score = \frac{\sum_k^N T_g(k) log[l_g(k)/l_f(k)]}{\sum_k^N T_g(k)}, \tag{5.3}$$

where $T_g(k)$ is the number of occurrences of bigram type k (e.g., $++L|+-L$) in the direction giver's speech segments, N is the number of possible bigrams (in this case, $N = (2 \times 2 \times 2 \times 3)^2 = 196$), and $l_g(k)/l_f(k)$ is the bigram likelihood estimates for direction giver's/direction follower's model. The likelihood estimates for a model m, are calculated from the speech data using

$$l_m(k) = \frac{T_m(k)}{\sum_{n=1}^N T_m(n)}. \tag{5.4}$$

5.3.3 Idiolect similarity measures

An idiolect is the language or speech of an individual during a particular period of his or her life. An individual's idiolectal style contains speaker identity information, and these traits could be expected to have an influence on the quality of linguistic information transmission, i.e., a speaker's use of slang, affected pronunciation, an accent, a lisp or a non-standard dialect would intuitively seem to make it more difficult for a listener who is not familiar with those particular idiolectal features to understand the speaker, and thus would impair transmission efficiency. For example, in the corpus, when describing the same section of the route some direction givers say "you want to go round the outside of that (forest)", while other direction givers say "a curve almost a half 'u' shape".

Language style matching [85]

In [85], researchers developed a method to calculate the similarity in dyads' use of function words, which they called language style matching (LSM). LSM calculates the frequency of occurrence of nine sorts of function words in a dialogue, and uses the following equation to calculate a final stylistic similarity

$$LSM = \frac{1}{N} \sum_{i=1}^{N} (1 - \frac{|A_i - B_i|}{A_i + B_i + 0.0001}) \tag{5.5}$$

where N represents the number of function word categories used for LSM (in this case $N = 9$); A_i and B_i represent the percentage of function words belonging to

Table 5.1: Word Categories Used for Calculating Language Style Matching[85]

Category	Examples
Personal pronouns	I, his, their
Impersonal pronouns	it, that, anything
Articles	a, an, the
Conjunctions	and, but, because
Prepositions	in, under, about
Auxiliary verbs	shall, be, was
High-frequency adverbs	very, rather, just
Negations	no, not, never
Quantifiers	much, few, lots

category i used by speaker A and speaker B during the whole dialogue, respectively, and 0.0001 is added to prevent empty sets. The function word categories used for calculating LSM are shown in Table 5.1. Note that instead of using an automatic word category counter such as Linguistic Inquiry and Word Count (LIWC), This study uses the manually annotated part-of-speech tagging provided in [131] for LSM calculation. It is believed that the manually annotated labels provide higher statistical accuracy when counting function words since it allows us to distinguish which part of speech a word is being used as. For example, the word "that" can be used as a definite article: "That man is my father", or as a conjunction: "He told me that he would be late". Note that different from other similarity measures used in this study, LSM was first proposed for social relationship study[134], and to our knowledge has never been used in any automatic speaker recognition systems.

Bigram model of part-of-speech (POS) tags

Textual bigram models are a common method of modeling text data [51]. As HCRC map task corpus only has an average of about one thousand words of each direction giver's speech, in this study, instead of actual words in the original paper, part-of-speech (POS) tags was used to build bigram models for each speaker. The POS tags used in this study are: verbs, nouns, adjectives, adverbs, auxiliary verbs, determiners, pronouns, prepositions, conjunctions and interjections. The similarity measure used in Section 5.3.2 (Eq. 5.3) is also used here.

5.4 Experiment

5.4.1 Experimental setup

For all of the models reported in this study, I used generalized linear mixed models to model the data with MFCCs ivector similarity, prosodic dynamic similarity, language style matching similarity, and part-of-speech bigram similarity as fixed effects, and group ID and map ID[1] as random effects. We began by building a base model with familiarity conditions, eye-contact conditions, and gender information as the fixed factor and then respectively added MFCCs ivector similarity, prosodic dynamic similarity, language style matching similarity, and part-of-speech bigram similarity to the base model as fixed effects. Model improvement was assessed by the likelihood ratio tests, which uses a chi-square test to examine whether likelihood improvement is significant with the additional predictor. If the added predictor significantly improved the model fit, it will be considered that the predictor accounted for a significant amount of variation in the dependent variable (i.e., path deviation). Functions provided by Matlab were used for these statistical significance analysis, and Poisson distribution was set to be the distribution of the dependent variable.

For all the similarity measures, the silent segments of the recordings were removed manually. For the MFCC based similarity measure, 39 dimensional MFCCs (with Δ and $\Delta\Delta$ values) were extracted every 10 ms using a 25 ms analysis window. For prosodic similarity measure, pitch and intensity contours were extracted using PRAAT [43] every 10 ms using a 25 ms analysis window. Syllable boundaries were extracted using Julius[135]. After extraction, the log values of every data point were calculated, and then linear interpolation was used to fill in the zero values. For bigram models, after calculating the appearance probability of every unigram and bigram, 0.001 was used to fill the zero-probability items. The bigram model was then renormalized so that its probability sum was 1. Also note that due to an experimental error during recording (restarting from the middle of the data) and extremely noisy recording conditions, three of the dialogues (q3ec5, q3nc3, and q6ec2) were removed from analysis.

[1]There were totally sixteen groups and sixteen maps. Each group provided eight dialogues using eight different maps.

5.4.2 Results

Fig. 5.3 shows the histograms of the similarity measures used in this study. We can see that while prosodic dynamic bigram similarity, language style matching similarity, and POS bigram similarity have similar distributions, MFCCs ivector similarity seem to has a more intensive distribution. Table 5.2 shows the model fitting comparison results between base predictor and the proposed predictor. Comparison results revealed that adding MFCCs ivector similarity, prosodic bigram similarity, and POS bigram similarity as predictor significantly improved the model fits($\chi^2 = 36.809, p < 0.01$; $\chi^2 = 6.397, p < 0.05$; $\chi^2 = 33.814, p < 0.01$, respectively). However, adding language style matching similarity as predictor did not improved the model fit ($\chi^2 = 0.1298, p > 0.05$). The results suggest that MFCCs ivector similarity, prosodic bigram similarity, and POS bigram similarity significantly influenced the final path deviation of the map task dialogue. Interlocutors with similar voices, who used similar sentence structures, and who exhibited similar prosodic behaviours, tended to achieve higher levels of linguistic information transmission quality. It may be a litter surprise that be similar in MFCCs ivector also facilitate linguistic information transmission quality, because it is well known that the perception of one's own voice involves a mixture of air conduction and bone conduction, meaning that our perception of our recorded voice differs from our daily perception of our own voices. It implies that although the spectral envelopes are different between these two types of voices (i.e., recorded voice and daily heard voice), our MFCCs including other speech characteristics which we are familiar with (e.g., fundamental frequency). Compared to MFCCs ivector similarity and POS bigram similarity, prosodic bigram similarity showed a lower significant level in our experiment. Maybe it is because of that participants in the HCRC map task were all born in Scotland, and therefore shared similarity prosodic dynamic patterns. Finally, LSM similarity did not significantly influence the linguistic information transmission quality. As LSM has usually been used to investigate social interaction and has never been evaluated with automatic speaker recognition task, function words are less important in linguistic information transmission quality and lack of discriminative capacity to measure the similarity between two speakers are two possible explanation of this result.

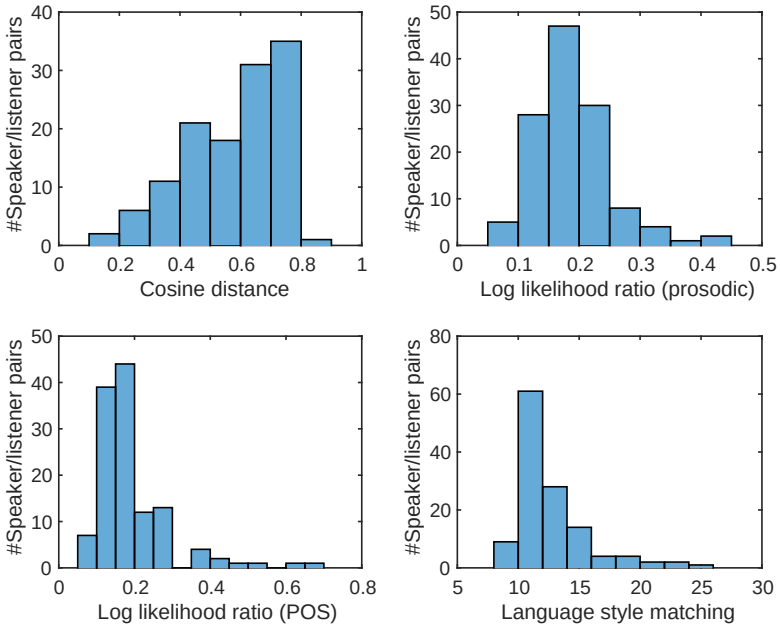

Figure 5.3: **Histograms of the similarity measures.** The upper left shows the histogram of MFCCs ivector similarity between direction giver and direction follower. The upper right shows the histogram of prosodic dynamic similarity between direction giver and direction follower. The lower left shows the histogram of LSM similarity between direction giver and direction follower. The lower right shows the histogram of POS bigram similarity between direction giver and direction follower.

Table 5.2: Results of likelihood ratio tests

Model	Predictor	χ^2	p
Null model	base		
Alternative model	base + ivector	36.809	< 0.01
Null model	base		
Alternative model	base + prosodic-bigram	6.397	< 0.05
Null model	base		
Alternative model	base + LSM	0.1298	> 0.05
Null model	base		
Alternative model	base + POS-bigram	33.814	< 0.01

5.5 Summary of this chapter

In this study relationships between the speaker-specific speech characteristics, such as segmental, prosodic and idiolectal similarity, and information transmission quality when speakers gave instructions using spoken language has been investigated. The results showed that interlocutors those with similar voices, who used similar sentence structures, and who exhibited similar prosodic behaviours, tended to achieve higher levels of linguistic information transmission accuracy. Also note that, the speech characteristics features for the participants were calculated using different dialogues, which implies that the prediction model could predict the accuracy of information transmission before the conversations had started.

Since the participants in the HCRC map task were all born in Scotland, this implies that they shared similar prosodic dynamic features and vocabularies. The results seems to support the idea that human language processing mechanisms are more sensitive than we generally believe. The facilitative effect of MFCCs ivector similarity consisted of our previous finding[130], which implies that MFCC feature including other speech characteristics which we are familiar with. For example, the fundamental frequency of the voice is the same between recorded voice and daily heard voice, and has already been demonstrated to have facilitative effect on implicit word recognition rate when words were read using the same fundamental frequency[129].

For future work, we plan to built a dialogue system based on the finding of this study that be similar with the speaker's voice could improve the linguistic information transmission quality. But there are still questions left, one of the most comments we heard is that it is very strange to hear somebody speaking using our own voice. Therefore, it is necessary to find out a compromised way to provide high quality information transmission while caring the listener's feeling.

Chapter 6

Conclusion and future work

6.1 Conclusion

In this dissertation I have described my investigations of the impact of speech characteristic similarity on information transmission efficiency in dialogue. Few previous studies in this area of research have attempted to measure the impact of quantized similarity in speech characteristics on communication among speakers, or have verified those effects using a spontaneous speech corpus. Moreover, the effects of segmental speech characteristics, such as formant information, similarity on information transmission have never been studied, because these characteristics are considered to contain physiological information, which is difficult to measure.

In my first study, I proposed a method to reduce clustering error during speaker diarization, which is the division of dialogues into speaker-specific clusters. The proposed method evaluated whether or not clusters were composed almost entirely of speech segments from only one speaker, using the statistical properties of inter-cluster similarity scores and intra-cluster similarity scores as measured in a so-called speaker space. Experimental results showed that the proposed cluster evaluation method could detect "ideal clusters" effectively, which improved the performance of the clustering algorithm by preventing over-merging, and hence diarization performance, compared to methods using conventional bottom-up clustering.

In my second study, a listening experiment was designed to investigate the impact of speech characteristic similarity on subtle prosodic information transmission ef-

ficiency at the segmental, prosodic, and lexical levels. Japanese right-branching (RB) vs. left- branching (LB) ambiguous sentences were used in the experiment in order to control lexical influence. Morphing technology and text-dependent objective similarity measures were introduced to control similarity levels. Participants were asked to finish a target identification task with RB vs. LB materials as targets. Their response times during the tasks, together with the proportion of their eye-fixation on different targets, were recorded for analysis. Results showed that similarity in speech characteristics, including similarity in segmental speech characteristics[1], which had not been investigated previously, has an apparently facilitative effect on subtle prosodic information transmission.

In the third study, speech characteristics similarity between subjects was proposed as a predictor of information transmission quality. The proposal was validated using a map task dialogue corpus. Typical speech characteristics similarity measures proposed for automatic speaker recognition based on their segmental features, prosodic features and lexical features had been selected. Likelihood ratio test of generalized mixed linear model was used to evaluate whether the selected similarity measures are effective predictors of the map task success. The results revealed an facilitative effect, with more successful in dialogues occurred between similar interlocutors than in dialogues occurred between dissimilar interlocutors.

In general, my investigations showed that similarity in speaker-specific speech characteristics between conversation partners facilitated information transmission efficiency. Comparison to familiarity of speech characteristics, which has been found facilitating linguistic information transmission by previous studies, speech characteristics similarity is more measurable which therefore makes self-similar voice a potential predictor and direction of high efficiency information transmission systems.

6.2 Future work

One possible direction for future research could be the exploration other speech characteristics and their associated properties which might have beneficial effects on information transmission efficiency. For example, since each speaker likely has

[1] In my research, this refers to the Euclidean distance between the MFCCs of aligned frames

his or her specific speech characteristics, being similar to one particular speaker may decrease a person's similarity to other speakers. In contrast, it is not hard to imagine that there may be speaker-specific speech characteristics which increase a speaker's information transmission efficiency in general for all listeners, such as the speech characteristics shared by news announcers (e.g., clear enunciation, steady pacing, high volume, etc.). Extracting these "general" speech characteristics which facilitate information transmission, and comparing their performance with that of "specific" speech characteristics (i.e., speech characteristics similar to those of the information receiver) is likely to be a focus of my future research.

Regarding practical applications of my research, on the one hand the investigations described in this dissertation suggests that the use of synthesized voices similar to those of system users can improve information transmission efficiency. On the other hand, it could be uncomfortable for customers to hear a voice identical to their own. Thus, we may need to balance high efficiency in conversation with user comfort, which could also be a big challenge.

In the future, just as smart phones have become indispensable in modern society, personal virtual assistants might become our closest confidants and most loyal supporters. These virtual assistants might serve not only as our private guides and secretaries, which is close to being achieved already, but also as our personal supervisors, counsellors and coaches who can help us process complex information, discuss with us the risks and benefits of various options and motivate us to take necessary action. Users might be able to customize their virtual personal assistant with various personalities, appearances, and speech characteristics. When facing tasks in which high efficiency information transmission is required, I believe the use of user-similar speech characteristics, at the prosodic, lexical, and even the segmental level, should be one of the available options.

Appendix A

Other 12 ambiguous material used in the ambiguous sentences interpretation experiment (Chapter 3)

Figure A.1: **The other 12 ambiguous items used in the experiment of Chapter 4.**

Bibliography

[1] Itakura, F., "Minimum prediction residual principle applied to speech recognition." Acoustics, Speech and Signal Processing, IEEE Transactions on, 23(1), (1975): 67-72.

[2] Lee, K.F., and Hon, H.W., "Large-vocabulary speaker-independent continuous speech recognition using HMM." In Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on (1988, April): 123-126.

[3] Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., and Kingsbury, B., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." Signal Processing Magazine, 29.6, (2012): 82-97.

[4] Reynolds, D.A., "Speaker identification and verification using Gaussian mixture speaker models." Speech communication, 17.1, (1995): 91-108.

[5] Parris, E.S., and Carey, M.J. "Language independent gender identification." In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference, (1996, May): 685-688.

[6] Minematsu, N., Sekiguchi, M., and Hirose, K., "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers." In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on (Vol. 1, pp. I-137), (2002, May).

[7] Pohjalainen, J., Rasanen, O., and Kadioglu, S., "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits." Computer Speech and Language, 29.1, (2015): 145-171.

[8] Scherer, S., Stratou, G., Gratch, J., and Morency, L. P., "Investigating voice quality as a speaker-independent indicator of depression and PTSD." In Interspeech, (2013): 847-851.

[9] Zhou, G., Hansen, J.H., and Kaiser, J.F., "Nonlinear feature based classification of speech under stress." Speech and Audio Processing, IEEE Transactions on, 9.3, (2001): 201-216.

[10] Bone, D., Black, M., Li, M., Metallinou, A., Lee, S., and Narayanan, S.S., "Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors." In INTERSPEECH, (2011, August): 3217-3220.

[11] Grimm, M., Kroschel, K., Mower, E., and Narayanan, S., "Primitives-based evaluation and estimation of emotions in speech." Speech Communication, 49.10, (2011, August): 787-800.

[12] Schuller, B., Rigoll, G., and Lang, M., "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference (Vol. 1, pp. I-577), (2004, May).

[13] Leggetter, C.J., and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models." Computer Speech and Language, 9.2, (1995): 171-185.

[14] Padmanabhan, M., Bahl, L.R., Nahamoo, D., and Picheny, M.A., "Speaker clustering and transformation for speaker adaptation in speech recognition systems." Speech and Audio Processing, IEEE Transactions on, 6.1, (1998): 71-77.

[15] Schuller, B., Stadermann, J., and Rigoll, G., "Affect-robust speech recognition by dynamic emotional adaptation." In Proc. speech prosody, (2006, May).

[16] Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G., "Combining frame and turn-level information for robust recognition of emotions within speech." In INTERSPEECH, (2007, August): 2249-2252.

[17] Mattheyses, W., and Verhelst, W., "Audiovisual speech synthesis: An overview of the state-of-the-art." Speech Communication, 66, (2015): 182-217.

[18] Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T., "Speaker adaptation for HMM-based speech synthesis system using MLLR." In The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, (1998).

[19] Schotz, S., "Towards synthesis of speaker age: A perceptual study with natural, synthesized and resynthesized stimuli." In Fonetik 2003 (Vol. 9, pp. 153-156). Dept. of Philosophy and Linguistics, Univ. Umea., (2003).

[20] Prahallad, K., Black, A.W., and Mosur, R., "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis." In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference (Vol. 1, pp. I-I), (2006, May).

[21] Yamagishi, J., and Kobayashi, T., "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training." IEICE TRANSACTIONS on Information and Systems, 90.2, (2007): 533-543.

[22] Schroder, M., "Emotional speech synthesis: a review." In INTERSPEECH, (2001, September): 561-564.

[23] Zovato, E., Pacchiotti, A., Quazza, S., and Sandri, S., "Towards emotional speech synthesis: A rule based approach." In Fifth ISCA Workshop on Speech Synthesis, (2004).

[24] Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M., and Narayanan, S. "Constructing emotional speech synthesizers with limited speech database." In Proc. ICSLP (Vol. 2, (2004, October): 1185-1188.

[25] Reynolds, D.A., "Channel robust speaker verification via feature mapping." Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference , 2003.

[26] Kitamura, T., "Acoustic analysis of imitated voice produced by a professional impersonator." INTERSPEECH, 2008.

[27] Kinnunen, T., and Li, H., "An overview of text-independent speaker recognition: From features to supervectors." Speech communication, 52.1, (2010): 12-40.

[28] Shriberg, E., "Higher-level features in speaker recognition." Speaker Classification I. Springer Berlin Heidelberg, (2007): 241-259.

[29] Davis, S., and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE Trans. Acoustics, Speech, Signal Process. 28.4, (1980): 357-366.

[30] Stevens, S.S., Volkmann, J., and Newman, E.B., "A scale for the measurement of the psychological magnitude pitch." The Journal of the Acoustical Society of America 8.3 (1937): 185-190.

[31] Chiba, T., and Kajiyama, M., "The Vowel: Its Nature and Structure." Tokyo: Tokyo-Kaiseikan Pub. Co., Ltd., (1942)

[32] Huang, X., Acero, A., and Hon, H. W., "Spoken Language Processing: a Guide to Theory." Algorithm, and System Development. Prentice-Hall, New Jersey, (2001).

[33] Hermansky, H., "Perceptual linear prediction (PLP) analysis for speech." J. Acoust. Soc. Amer. 87, (1990): 1738-1752

[34] Lu, X., and Dang, J., "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification." Speech Comm. 50.4, (2007): 312-322.

[35] Magrin-Chagnolleau, I., Durou, G., and Bimbot, F., "Application of time frequency principal component analysis to text-independent speaker identification." IEEE Trans. Speech Audio Process. 10.6, (2002): 371-378.

[36] Malayath, N., Hermansky, H., Kajarekar, S., and Yegnanarayana, B., "Data-driven temporal filters and alternatives to GMM in speaker verification." Digital Signal Process. 10.1, (2000): 55-64.

[37] Kinnunen, T., "Spectral Features for Automatic Text-Independent Speaker Recognition." Licentiate's Thesis, University of Joensuu, Department of Computer Science, Joensuu, Finland, (2004).

[38] Furui, S., "Cepstral analysis technique for automatic speaker verification." IEEE Trans. Acoustics, Speech Signal Process. 29.2, (1981): 254-272.

[39] Farrs, M., Hernando, J., and Ejarque, P., "Jitter and shimmer measurements for speaker recognition." INTERSPEECH, (2007).

[40] Friedland, G., Vinyals, O., Huang, Y., and Muller, C., "Prosodic and other long-term features for speaker diarization." IEEE Transactions on Audio, Speech, and Language Processing, 17.5, (2009): 985-993.

[41] Titze, I., "Principles of Voice Production." Prentice Hall, Englewood Cliffs, (1994)

[42] Parris, E. S., and Michael J. C., "Language independent gender identification." Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference, (1996).

[43] Praat software website: http://www.fon.hum.uva.nl/praat/.

[44] Eskelinen-Ronka, P., and Niemi-Laitinen, T., "Testing voice quality parameters in speaker recognition." In Proc. The 14th Int. Congress on Phonetic Sciences (ICPhS 1999), San Francisco, California, USA, (1999): 149-152.

[45] Leena, M., and Yegnanarayana. B., "Extraction and representation of prosodic features for language and speaker recognition." Speech communication, 50.10, (2008): 782-796.

[46] Sonmez, K., Shriberg, E., Heck, L., and Weintraub, M., "Modeling Dynamic Prosodic Variation for Speaker Verification." In: Mannell, R.H., Robert-Ribes, J. (eds.) Proc. ICSLP. vol. 7, Australian Speech Science and Technology Association, Sydney, (1998): 3189-3192.

[47] Adami, A. G., Mihaescu, R., Reynolds, D. A., and Godfrey, J.J., "Modeling Prosodic Dynamics for Speaker Recognition." In: Proc. ICASSP. Hong Kong, 4, (2003): 788-791.

[48] Chen, S.H., and Wang, H.C., "Improvement of Speaker Recognition by Combining Residual and Prosodic Features with Acoustic Features." In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, (2004).

[49] Chen, J., Dai, B., and Sun, J., "Prosodic Features Based on Wavelet Analysis for Speaker Verification." In: Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech), Lisbon, Portugal, (2005): 3093-3096.

[50] Chen, Z.H., Liao, Y.F.L., and Juang, Y.T., "Eigen-Prosody Analysis for Robust Speaker Recognition under Mismatch Handset Environment." In: Proceedings of the International Conference of Spoken Language Processing (ICSLP 04), Jeju Island, South Korea, (2004)

[51] Doddington, G., "Speaker recognition based on idiolectal differences between speakers." In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, September, (2001): 2521-2524.

[52] Gamon, M., "Linguistic correlates of style: Authorship classification with deep linguistic analysis features." In Proceedings of the 20th International Conference on Computational Linguistics, (2004): 611-617.

[53] Park, A., Hazen, T.J., "ASR Dependent Techniques for Speaker Identification." In: Hansen, J.H.L., Pellom, B. (eds.) Proc. ICSLP, Denver, (2002): 1337-1340.

[54] Boakye, K., and Peskin, B., "Text-Constrained Speaker Recognition on a Text-Independent Task." In: Proceedings Odyssey-04 Speaker and Language Recognition Workshop, Toledo, Spain, (2004)

[55] Peng, F., Shuurmans, D., and Wang, S., "Augmenting naive Bayes classifiers with statistical language models." Information Retrieval Journal, 7.1, (2004): 317-345.

[56] Sanderson, C., and Guenter, S., "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation." In Proceedings of the International Conference on Empirical Methods in Natural Language Engineering, (2006): 482-491.

[57] Hirst, G., and Olga, F. "Bigrams of syntactic labels for authorship discrimination of short texts." Literary and Linguistic Computing, 22.4, (2007): 405-417.

[58] Hancock, J.T., Curry, L., Goorha, S., and Woodworth, M.T., "On lying and being lied to: A linguistic analysis of deception." Discourse Processes, 45, (2007): 1-13.

[59] Slatcher, R.B., Chung, C.K., Pennebaker, J.W., and Stone, L.D., "Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates." Journal of Research in Personality, 41, (2007): 63-75.

[60] Pickering, M.J., and Ferreira, V.S., "Structural priming: A critical review." Psychological bulletin, 134.3, (2008): 427-459.

[61] Pickering, M.J., and Garrod, S., "An integrated theory of language production and comprehension." Behavioral and Brain Sciences, 36.4, (2013): 329-347.

[62] Kintsch, W., and van Dijk, T.A., "Toward a model of text comprehension and production." Psychological Review, 85, (1978): 363-394.

[63] Zwaan, R.A., "The immersed experiencer: Toward an embodied theory of language comprehension." Psychology of learning and motivation, 44, (2003): 35-62.

[64] Ericsson, K.A., and Kintsch, W., " ." Psychological Review, 102, (1995): 211-245.

[65] Meyer, D.E., and Schvaneveldt, R.W., "Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations." Journal of Experimental Psychology. 90, (1971): 227-234.

[66] Weiner, E.J., and Labov, W., "Constraints on the agentless passive." Journal of Linguistics, 191, (1983): 29-58.

[67] Branigan, H.P., Pickering, M.J., and McLean, J.F., "Priming prepositional-phrase attachment during comprehension." Journal of Experimental Psychology: Learning, Memory, and Cognition, 31.3, (2005): 468-481.

[68] Bartlett, F.C., and Burt, C., "Remembering: A study in experimental and social psychology." British Journal of Educational Psychology, 3.2, (1933): 187-192.

[69] Rumelhart, D., "Schemata: The building blocks of cognition." In. R. Spiro, B. Bruce and W. Brewer (eds.) Theoretical issues in reading comprehension, (1980).

[70] Morrow, D.G., and Clark, H.H., "Interpreting words in spatial descriptions." Language and Cognitive Processes, 3, (2003): 275-291.

[71] Buchel. C., Price, C., and Friston, K., "A multimodal language region in the ventral visual pathway." Nature, 392, (1998): 274-277.

[72] Zwaan, R.A., Stanfield, R.A., and Yaxley, R.H., "Do language comprehenders routinely represent the shapes of objects?" Psychological Science, 13, (2002): 168-171.

[73] DeLong, K.A., Urbach, T.P., and Kutas, M., "Probabilistic word pre-activation during language comprehension inferred from electrical brain activity." Nature Neuroscience, 8.8, (2005):1117-1121.

[74] Pickering, M.J, Garrod, S., "Toward a mechanistic psychology of dialogue." Behavioral and Brain Sciences, 27, (2004): 169-225.

[75] Martin, C.S., Mullennix, J.W., Pisoni, D.B., and Summers, W.V., "Effects of talker variability on recall of spoken word lists." Journal of Experimental Psychology: Learning, Memory, and Cognition, 15.4, (1989): 676-684.

[76] Goldinger, S.D., Pisoni, D.B., and Logan, J.S., "On the nature of talker variability effects on recall of spoken word lists." Journal of Experimental Psychology: Learning, Memory, and Cognition, 17.1, (1991): 152-162.

[77] Van Lancker, D., and Canter, J., "Impairment of voice and face discrimination in patients with hemispheric damage." Brain and Cognition, 1, (1982): 185-195.

[78] Von Kriegstein, K., Eger, E., Kleinschmidt, A., and Giraud, A.L., "Modulation of neural responses to speech by directing attention to voices or verbal content." Cognitive Brain Research, 17, (2003): 48-55.

[79] Van Berkum, J.J., Van den Brink, D., Tesink, C.M., Kos, M., and Hagoort, P."The neural integration of speaker and message." Journal of cognitive neuroscience, 20.4, (2008): 580-591.

[80] Levitan, R., and Hirschberg, J., "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." Interspeech, (2011).

[81] Levitan, R., Gravano, A., and Hirschberg, J., "Entrainment in speech preceding backchannels." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2. Association for Computational Linguistics, (2011).

[82] Nenkova, A., Gravano, A., and Hirschberg. J., "High frequency word entrainment in spoken dialogue." Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers. Association for Computational Linguistics, (2008).

[83] Reitter, D., and Moore, J.D., "Predicting success in dialogue." In Proc. 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic, (2007): 808-815.

[84] Kinzler, K.D., Dupoux, E., and Spelke, E.S., "The native language of social cognition." Proceedings of the National Academy of Sciences 104.30, (2007): 12577-12580.

[85] Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., and Pennebaker, J.W., "Language style matching predicts relationship initiation and stability." Psychological science 22.1, (2011): 39-44.

[86] Tranter, S.E., and Reynolds, D.A., "An overview of automatic speaker diarization systems." IEEE Transactions on audio, speech, and language processing, 14.5, (2006) 1557-1565.

[87] Wooters, C., and Huijbregts, M., "The ICSI RT07s speaker diarization system." Multimodal Technologies for Perception of Humans, (2008): 509-519.

[88] Anguera, X., and Bonastre, J.F., "Fast speaker diarization based on binary keys." In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference, (2011): 4428-4431.

[89] Reynolds, D.A., and Torres-Carrasquillo, P.A., "The MIT Lincoln laboratory RT-04F diarization systems: Applications to broadcast news and telephone conversations." Proc. Fall 2004 RT-04 Workshop, (2004).

[90] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P., "Front-end factor analysis for speaker verification." IEEE Transactions on Audio, Speech, and Language Processing, 19.4, (2011): 788-798

[91] Tsai, W.H., Cheng, S., and Wang, H., "Speaker clustering of speech utterances using a voice characteristic reference space." In INTERSPEECH, (Oct. 2004).

[92] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P., "Joint factor analysis versus eigen channels in speaker recognition." IEEE Transactions on Audio, Speech, and Language Processing, 15.4, (2007): 1435-1447.

[93] Vijayasenan, D., Valente, F., and Bourlard, H., "An information theoretic approach to speaker diarization of meeting data." IEEE Transactions on Audio, Speech, and Language Processing, 17.7, (2009): 1382-1393.

[94] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., and Kronenthal, M., "The AMI meeting corpus." In Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, 88, (Sept. 2005).

[95] Larcher, A., Bonastre, J.F., Fauve, B.G., Lee, K.A., Levy, C., Li, H., and Parfait, J.Y., "ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition." In Interspeech (2013): 2768-2772.

[96] Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D.A., and Glass, J.R., "Exploiting Intra-Conversation Variability for Speaker Diarization." In interspeech, (2011): 945-948.

[97] http://www.nist.gov/itl/

[98] Liu, D., and Kubala, F., "Online speaker clustering." In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference, (2003): 333-336.

[99] Rand, W.M., "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association, 66.336, (1971): 846-850.

[100] Anderson, N.J., and Cheng, X., "Exploring second language reading: Issues and strategies." Boston, MA: Heinle and Heinle, (1999): 53-56.

[101] Slobin, D.I., "Grammatical transformations and sentence comprehension in childhood and adulthood." Journal of verbal learning and verbal behavior, 5.3, (1966): 219-227.

[102] Frazier, L., and Rayner, K., "Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences." Cognitive psychology, 14.2, (1982): 178-210.

[103] Nassaji, H., "Schema Theory and Knowledge Based Processes in Second Language Reading Comprehension: A Need for Alternative Perspectives." Language learning, 52.2, (2007): 439-481.

[104] Leeser, M.J., "Learner based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory." Language Learning, 57.2, (2007): 229-270.

[105] Lee, S.K., "Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form." Language learning, 57.1, (2007): 87-118.

[106] Erten, I.H., and Razi, S., "The Effects of Cultural Familiarity on Reading Comprehension." Reading in a Foreign Language, 21.1, (2009): 60-77.

[107] Adank, P., Evans, B.G., Stuart-Smith, J., and Scott, S.K., "Comprehension of familiar and unfamiliar native accents under adverse listening conditions." Journal of Experimental Psychology: Human Perception and Performance, 35.2, (2009): 520-529.

[108] Major, R.C., Fitzmaurice, S.F., Bunta, F., and Balasubramanian, C., "The effects of nonnative accents on listening comprehension: Implications for ESL assessment." TESOL quarterly, (2002): 173-190.

[109] Hirose. Y., "Cognitive mechanisms for sentence comprehension speaker's intention and hearer's comprehension: A latent function of lexical accent in syntax." Cognitive Science 13.3, (2006): 428-442.

[110] Barton, J.J., Radcliffe, N., Cherkasova, M.V., Edelman, J., and Intriligator, J.M., "Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations." Perception, 35.8, (2006): 1089-1105.

[111] Valentine, T., Darling, S., and Donnelly, M., "Why are average faces attractive? The effect of view and averageness on the attractiveness of female faces." Psychonomic Bulletin and Review, 11.3, (2004): 482-487.

[112] Kawahara, H., Takahashi, T., Morise, M., and Banno, H., "Development of exploratory research tools based on TANDEM-STRAIGHT." In Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, (2009): 111-120.

[113] Skuk, V.G., and Schweinberger, S.R., "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender." Journal of Speech, Language, and Hearing Research, 57.1, (2014): 285-296.

[114] Zaske, R., Schweinberger, S. R., and Kawahara, H., "Voice aftereffects of adaptation to speaker identity." Hearing Research, 268.1, (2010):38-45.

[115] Sakoe, H., and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition." IEEE transactions on acoustics, speech, and signal processing, 26.1, (1978): 43-49.

[116] Hermes, D.J., "Measuring the perceptual similarity of pitch contours." Journal of Speech, Language, and Hearing Research, 41.1, (1998): 73-82.

[117] Kawahara, H., de Cheveigne, A., Banno, H., Takahashi, T., and Irino, T., "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT." In Interspeech, (2005): 537-540.

[118] Mattys, S.L., and Liss, J.M. "On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality." Attention, Perception, and Psychophysics, 70(7), (2008):1235-1242. ISO 690

[119] Levi, S.V., "Talker familiarity and spoken word recognition in school-age children." Journal of child language 42(04) (2015): 843-872.

[120] Goggin, J.P., Thompson, C.P., Strube, G., and Simental, L.R., "The role of language familiarity in voice identification.", Memory and Cognition, 19, (1991): 448-458.

[121] Thompson, C.P., "A language effect in voice identification." Applied Cognitive Psychology, 1, (1987): 121-131.

[122] Winters, S.J., Levi, S.V., and Pisoni, D.B., "Identification and discrimination of bilingual talkers across languages." Journal of the Acoustical Society of America, 123, (2008): 4524-4538.

[123] Borrie, S.A., McAuliffe, M.J., Liss, J.M., O'Beirne, G.A., and Anderson, T.J., "The role of linguistic and indexical information in improved recognition of dysarthric speech." The Journal of the Acoustical Society of America, 133(1), (2013): 474-482.

[124] Nathan, L., Wells, B., and Donlan, C., "Children's comprehension of unfamiliar regional accents: a preliminary investigation." Journal of child language, 25.2, (1998): 343-365.

[125] Adank, P., and Janse, E. "Comprehension of a novel accent by young and older listeners." Psychology and aging, 25.3, (2010): 736-740.

[126] Adank, P., Hagoort, P., and Bekkering, H. "Imitation improves language comprehension." Psychological Science, 21012, (2010): 1903-1909

[127] Nygaard, L.C., Sommers, M.S., and Pisoni, D.B., "Speech perception as a talker-contingent process." Psychological Science, 5.1, (1994): 42-46.

[128] Nygaard, L.C., and Pisoni, D.B., "Talker-specific learning in speech perception." Attention, Perception, and Psychophysics, 60(3), (1998): 355-376.

[129] Church, B.A., and Schacter, D.L.. "Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency." Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(3), (1994): 521-533.

[130] Chen, B., Kitaoka, N., and Takeda, K., "Impact of acoustic similarity on efficiency of verbal information transmission via subtle prosodic cues." EURASIP Journal on Audio, Speech, and Music Processing, 2016.1, (2016).

[131] Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., and Sotillo, C., "The HCRC map task corpus." Language and speech, 34.4, (1991): 351-366.

[132] Greenberg, C.S., Bans, D., Doddington, G.R., Garcia-Romero, D., Godfrey, J.J., Kinnunen, T., ... and Reynolds, D.A.. "The NIST 2014 speaker recognition i-vector machine learning challenge." In Odyssey: The Speaker and Language Recognition Workshop (pp. (2014): 224-230.

[133] Adami, A.G., Mihaescu, R., Reynolds, D.A., and Godfrey, J.J., "Modeling prosodic dynamics for speaker recognition." In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference, (2003)

[134] Niederhoffer, K.G., and Pennebaker, J.W., "Linguistic style matching in social interaction." Journal of Language and Social Psychology, 21(4), (2002): 337-360.

[135] Lee, A., Kawahara, T., and Shikano, K., "Julius — an open source real-time large vocabulary recognition engine." In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), (2001): 1691-1694.

List of Publications

Journal Papers

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Modified bottom-up clustering based on evaluation of speech segment cluster for improved speaker diarization." IEICE TRANSACTIONS on Information and Systems D 97.3, (2014): 540-547. (in Japanese)

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Impact of acoustic similarity on efficiency of verbal information transmission via subtle prosodic cues." EURASIP Journal on Speech, Audio, and Music Processing, (2016), 2016:19.

International Conference Proceedings

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Effect of speaking rate and speech complexity on transmission quality during driving navigation task." Seventh Biennial Workshop on Digital Signal Processing for In-Vehicle Systems (DSP in Vehicles 2015), Berkeley, USA, Oct., 2015.

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Relationship between Speaker/Listener Similarity and Information Transmission Quality in Speech Communication", APSIPA 2015, Hong Kong, China, Dec., 2015

Domestic Conference Proceedings

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Cluster Purity Estimation for

Highly Accurate Speaker Diarization." Acoustical Society of Japan, 1-1-14, (4 pages), Sept., 2012. (in Japanese)

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Cluster selection for Highly Accurate Speaker Diarization." Symposium on Spoken Linguistics, (6 pages), Dec. 2012. (in Japanese)

Bohan Chen, Norihide Kitaoka, Mihoko Otake, Kazuya Takeda, "Evaluation of speaker engagement using turn-taking behavior entropy." Technical Report of IEICE, SP/SLP 2015-52, pp. 13-17, Jun., 2014.

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Correlation between information transmission quality and interlocutor similarity." Acoustical Society of Japan, 2-Q-20, (4 pages), Sept., 2014. (in Japanese)

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Rational controls of speech characteristics and their influences on transmission efficiency during information transmission by speech." Acoustical Society of Japan, 1-R-20, (4 pages), Mar., 2015.

Bohan Chen, Norihide Kitaoka, Mihoko Otake, Kazuya Takeda, "Statistical turn-taking modeling for speaker engagement level estimation based on entropy." Acoustical Society of Japan, 1-Q-35, (4 pages), Sep., 2015.

Bohan Chen, Norihide Kitaoka, Kazuya Takeda, "Difference of prosodic information transmission efficiency casued by verbally meaningless acoustic difference : An experimental study." Acoustical Society of Japan, 2-Q-32, (4 pages), Mar., 2016.

Ryota Nishimura, **Bohan Chen**, Norihide Kitaoka, "OOV registration to language model for speech recognition." Acoustical Society of Japan, 1-Q-23, (4 pages), Mar., 2018. (accepted)