

Intelligibility of English mosaic speech: Influence of manipulating mosaic block duration

サンティ

<https://hdl.handle.net/2324/4475136>

出版情報 : Kyushu University, 2020, 博士 (芸術工学) , 課程博士
バージョン :
権利関係 :



KYUSHU UNIVERSITY

Graduate School of Design

Department of Human Science

Human Science International Course – Doctoral Program

**Intelligibility of English mosaic speech:
Influence of manipulating mosaic block duration**

**英語モザイクスピーチの明瞭度：
モザイクブロック幅の操作による影響**

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF HUMAN SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF KYUSHU UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Principal Advisor:

Dr. Gerard B. Remijn

Santi

3DS16021E

February 2021

Acknowledgements

I would like to big thank Prof. Emeritus Yoshitaka Nakajima and Prof. Dr. Gerard B. Remijn for their kindnesses continuous support of my doctoral study and research professionally and patiently, for encouragement and professional knowledge supporting me in finishing this dissertation. I would like also to thank Prof. Dr. Kazuo Ueda for his advice and comments in the numerous seminars. Furthermore, I thank also all the students from the Professor Nakajima Laboratory (October 2016 – March 2020), from the Professor Remijn Laboratory, and from the Professor Ueda Laboratory, especially for Masamichi Hinoshita, Takahiro Ishikawa, Hikaru Eguchi, and Aiko Nakata for coding the mosaic speech program, Takuya Kishida for technical assistance, Shimeng Liu, Yesaya Tommy Paulus, João Paulo Cabral for their support in data collection, Alexandra Wolf for her comments on an initial version of my article manuscript, and other students for their friendship and to help me with their valuable data for my pilot experiments.

Moreover, the most important thanks to my family, especially to my husband Samsul, my mother Hanisang, my father Saleh, my lovely daughter Sasya, and my brothers Muliadi and Sandip Ghimire, for their loves and supports in each day.

Lastly, I would like to thank the staff of Kyushu University for their kind and professional attention, and the Japan Student Services Organization (JASSO; October 2016 – March 2017) and STMIK Dipanegara Makassar for supporting my studies.

Abstract

Mosaic speech is degraded speech that is segmented into time \times frequency blocks. Earlier research with Japanese mosaic speech has shown that the intelligibility of mosaic speech was almost perfect for mosaic block durations (MBD) of 20 and 40 ms. The first objective of the present research was to investigate the intelligibility of English mosaic speech, and whether its intelligibility would vary if it was compressed, preserved, or stretched in time. The second objective was to investigate whether the effects of compressing, preserving or stretching mosaic speech would be similar among listeners with different language backgrounds. To achieve these objectives, two experiments were conducted. The preliminary experiment was conducted first with Indonesian listeners ($n= 20$) followed by the main experiment with native-English ($n= 19$), Indonesian ($n= 19$), and Chinese ($n= 20$) listeners. In the experiments, English mosaic words were presented to the participants, and they typed what they had heard. The intelligibility of English mosaic speech (individual words) was obtained by counting the number of correct words given by the participants.

For the first objective, from the two experiments conducted, it was found that the listeners from the three language groups (native-English, Indonesian, and Chinese) showed the same trends in intelligibility scores: English mosaic speech was most intelligible when the OMBDs were preserved or stretched into 20- or 40-ms MBDs. Intelligibility decreased when the OMBDs were compressed, or stretched into MBDs of 80 ms or longer. The results seem to agree with the results of earlier research, which showed that mosaic speech is most intelligible in the segment duration range of 20 to 40 ms. When the segment is longer, the intelligibility becomes lower, but the results thus show that even rather long 40-ms segments can be processed as intelligible speech.

The second objective of this thesis research was to investigate whether the effects of compressing, preserving, or stretching mosaic speech varies among listeners with different language backgrounds. The results showed that the intelligibility was relatively high for stimuli with preserved OMBDs of 20 ms and 40 ms for all language groups, and also for stimuli with an MBD of 40 ms after stretching the OMBD of 20 ms, but only for the native-English group. The OMBD was manipulated by compressing or stretching it without changing its linguistic information. However, the speed of speech changed and this caused the intelligibility to change as well. Both non-native listener groups showed the same trend regarding the speed of speech, that is, the intelligibility was highest for the preserved speech. However, the native-English listeners obtained the highest intelligibility scores for preserved speech or slightly slower speech, but this happened only when the OMBD of 20 ms was stretched into a 40-ms MBD. Thus, this thesis research suggests that presenting the same acoustic information in any temporal segment does not guarantee that the intelligibility will be preserved, but the temporal segment duration plays the most important role to determine intelligibility in mosaic speech perception. In other words, the intelligibility was affected by mosaic block duration (MBD). Regarding the stimuli with the same (preserved/stretched) MBDs among both OMBDs, i.e. 40, 80, or 160 ms, the intelligibility did not change significantly even when the amount of information must have changed for the native-English and the Indonesian listeners, but not for the Chinese listeners. Thus, this thesis research suggests that the intelligibility was not affected by OMBD when it was preserved/stretched in the range of 40-160 ms.

In general, the results of the thesis research suggest that humans can extract new information from individual speech segments of about 40 ms, but that there is a limit to the amount of linguistic information that can be conveyed within a block of about 40 ms or below.

Research Output

The following Chapters of this thesis research have been presented in the following conference proceedings or academic peer-reviewed journals.

1. Chapter 2 was presented in an oral session of the 35th Annual Meeting of the International Society for Psychophysics, Fechner Day 2019, Antalya, Turkey, 30 October–2 November 2017.

Santi, S., Nakajima, Y., Ueda, K., Remijn, G.B. (2019). Effects of compressing or stretching mosaic block duration on intelligibility of English mosaic speech. In *Proceedings of the 35th Annual Meeting of the International Society for Psychophysics, Fechner Day 2019*, p. 35.

2. Chapter 3 was published in:

Applied Science of Multidisciplinary Digital Publishing Institute (MDPI).

Santi, Nakajima, Y., Ueda, K., Remijn, G. B. (2020). Intelligibility of English Mosaic Speech: Comparison between Native and Non-Native Speakers of English. *Appl. Sci.* *10*, 6920. 2020/10/02. doi:10.3390/app10196920.

Table of Contents

Acknowledgements	II
Abstract	III
Research Output	V
Table of Contents	VI
List of Figures	IX
List of Tables	XI
List of Abbreviations	XII
Chapter 1 -Introduction	1
1.1 Accoustic Properties of Speech	2
1.2 Speech Perception and Speech Intelligibility.....	4
1.3 Characteristics of English Speech and Its Intelligibility.....	5
1.4 Speech Intelligibility in Environtmental Noise and Reverberation	6
1.5 Intelligibility of Temporally Manipulated or Segmented Speech.....	8
1.6 Neuroscientific Findings Related to Temporal Aspects of Speech Processing.....	9
1.7 Mosaic Speech	10
1.8 Compressed and Stretched Speech	12
1.9 General Purpose	13
1.10 Structure of the Dissertation	15
Chapter 2 -Preliminary Experiment: Effects of Compressing or Stretching Mosaic Block Duration on English Speech Intelligibility	17
2.1 Purpose.....	17
2.2 Method	17
2.2.1 Participants.....	17
2.2.2 Equipment	18

2.2.3 Stimuli.....	19
2.2.4 Procedures.....	25
2.2.5 Statistical Analysis.....	26
2.3 Results	28
2.3.1 The Effect of Compressing or Stretching the Original Mosaic Block Duration (OMBD).....	29
2.3.2 Intelligibility Comparisons between Stimuli with the Same MBDs.....	31
2.4 Discussion	32
Chapter 3 -Main Experiment: Speech Intelligibility Comparison between Native and Non-Native Speakers of English	35
3.2 Purpose	35
3.3 Method	36
3.3.1 Participants.....	36
3.3.2 Equipment.....	37
3.3.3 Stimuli.....	38
3.3.4 Procedures.....	41
3.3.5 Statistical Analysis.....	42
3.4 Results.....	42
3.4.1 Intelligibility Comparisons between Original Speech and Mosaic Speech	42
3.4.2 The Effect of Compressing or Stretching the Original Mosaic Block Duration (OMBD).....	44
3.4.3 Intelligibility Comparisons between Stimuli with the Same MBDs within Each Language Group.....	48
3.4.4 Intelligibility Comparisons between the Language Groups.....	50
3.4.5 Intelligibility of Phonemes of the Initial Consonant of Each English Word	52
3.5 Discussion	58
Chapter 4 - General Discussion and Conclusions	61

References	67
Appendix A. Informed consent and instruction of stimulus recording (preliminary experiment)	77
Appendix B. Stimulus recording procedure	78
Appendix C. Informed consent and instruction of listening experiment (Preliminary experiment)	79
Appendix D. Statistical analysis of preliminary experiment data	83
Appendix E. Sound pressure level (SPL) of original speech sounds (Fast peak) Preliminary experiment)	84
Appendix F. Sound pressure level (SPL) of original speech sounds (Fast peak) Main experiment)	87
Appendix G. Informed consent and instruction of listening experiment (Main experiment)	88
Appendix H. Statistical analysis of Main experiment data	90

List of Figures

Figure 1.1. The transmission of the linguistic codes from the speaker to the listener (from: de Saussure, 1966)	2
Figure 1.2. An original visual image (a, left) changed into mosaic image (a, right) of Fukuoka Tower, Fukuoka, Japan. A mosaic image can be created by averaging the color values or luminance grades within each block. An original speech (b, left) was changed into mosaic speech (b, right) by averaging the total amount of sound energy (as indicated by yellow and orange colors; taken from the Cool Edit 2000 software window)	12
Figure 2.1. Examples of the mosaic speech stimuli used throughout this thesis. (a) The waveform and (b) the spectrogram of the original speech for the word “mouse”, pronounced by a male native-English speaker. (c) An example of mosaic speech with an original mosaic block duration (OMBD) of 40 ms. Each individual block consisted of one mosaic block duration on the horizontal axis and one frequency band on the vertical axis. (d) An example of compressed mosaic speech with an OMBD of 40 ms compressed into a mosaic block duration (MBD) of 20 ms. (e) An example of stretched mosaic speech consisting of an OMBD of 40 ms stretched into an MBD of 80 ms	24
Figure 2.2. Results of the preliminary experiment. English word identification accuracy (intelligibility) for mosaic speech as a function of MBD after compressing/preserving/stretching (Indonesian, n = 20). The data for half-phase and whole-phase types are collapsed. Error bars indicate standard errors	29
Figure 3.1. Results of the main experiment. Word identification accuracy (intelligibility) for original speech for each language group (English, n = 19; Indonesian, n = 19; Chinese, n = 20). Error bars indicate standard error of means	42
Figure 3.2. Results of the main experiment. English word identification accuracy (intelligibility) for mosaic speech as functions of MBD after compressing/preserving/stretching (English, n = 19; Indonesian, n = 19; Chinese, n = 20). The data for the half-phase and the whole-phase types are collapsed. Error bars indicate standard errors	43
Figure 3.3. The intelligibility of phonemes of English words as functions of MBD after compressing/preserving/stretching the OMBD of 20 ms (A) and the OMBD of 40 ms (B) for the native-English group (n=19). H indicates the half-phase type and W indicates the whole-phase type of mosaicizing phase	54
Figure 3.4. The intelligibility of phonemes of English words as functions of MBD after compressing/preserving/stretching the OMBD of 20 ms (A) and the OMBD of 40 ms (B) for the Indonesian group (n=19). H indicates the half-phase type and W indicates the whole-phase type of mosaicizing phase.	56

Figure 3.5. The intelligibility of phonemes of English words as functions of MBD after compressing/preserving/stretching the OMBD of 20 ms (A) and the OMBD of 40 ms (B) for the Chinese group (n=20). H indicates the half-phase type and W indicates the whole-phase type of mosaicizing phase 57

List of Tables

Table 2.1. The 80 words used in the English mosaic speech preliminary experiment	20
Table 2.2. Mosaic speech block durations used in the preliminary experiment	23
Table 2.3. The assignment of stimulus type to the word groups for each participant, as used in the preliminary experiment	26
Table 2.4. Results of the preliminary experiment. Results of the Wilcoxon Signed-rank test for comparing the intelligibility between mosaicized stimuli with a half- and a whole-phase type of the OMBD of 20 and 40 ms	28
Table 2.5. Results of the preliminary experiment. Intelligibility comparisons between mosaic speech stimuli with a different MBD after compressing, preserving or stretching.....	31
Table 2.6. Results of the preliminary experiment. Multiple comparisons of intelligibility between stimuli with the same MBDs	32
Table 3.1. The 80 CVC-words used in the main experiment about English mosaic speech	40
Table 3.2. Results of the main experiment. Multiple comparisons of intelligibility between compressed, preserved, and stretched mosaic speech conditions within the OMBD of 20 or 40 ms for the native-English listeners (n=19)	45
Table 3.3. Results of the main experiment. Multiple comparisons of intelligibility between compressed, preserved, and stretched mosaic speech conditions within the OMBD of 20 or 40 ms for the Indonesian listeners (n=19)	46
Table 3.4. Results of the main experiment. Multiple comparisons of intelligibility between compressed, preserved, and stretched mosaic speech conditions within the OMBD of 20 or 40 ms for the Chinese listeners (n=20)	48
Table 3.5. Results of the main experiment. Multiple comparisons of intelligibility between stimuli with the same MBDs within each language group	50
Table 3.6. Results of the main experiment. Intelligibility comparisons between the language groups	51

List of Abbreviations

CET	College English Test
CV	Consonant Vowel
CVC	Consonant vowel consonant
dBA	Decibel-A weighting
EFL	English as a foreign language
ENL	English as a native language
ESL	English as a second language
F ₀	Fundamental frequency
Hz	Hertz
IELTS	International English Language Testing System
kHz	Kilohertz
MBD	Mosaic block duration
ms	Millisecond
OMBD	Original mosaic block duration
TFS	Temporal fine structure
TOEFL-iBT	Test of English as a Foreign Language-Internet-based Test
TOEFL-ITP	Test of English as a Foreign Language-Institutional Testing Program
TOEIC	Test of English for International Communication
VC	Vowel consonant

Chapter 1 - Introduction

In daily life, wherever humans live together, they build a system to communicate with each other and speech is a natural tool to make this communication (Carré et al., 2017; Denes & Pinson, 2015), by using a language (Feldman, 2019). Speech is an essential activity in human life, and a common and efficient form of communication by which humans can exchange their cultures, share experiences, ideas, and knowledge with others (Denes & Pinson, 2015). From a psycholinguistic viewpoint, speech can be defined as an individual act of willingness and intelligence to create a combination of language codes, so that humans can express their thoughts and feelings (de Saussure, 1966). The speech starts from a thought which originates in one person's brain (i.e., a speaker), and gets communicated to that of another (i.e., a listener) via a set of arbitrary associations between concepts and sound images that exist within a given language. This is known as “the speech circuit” (Figure 1.1; de Saussure, 1966). In detail, the circuit starts from the speaker’s brain (psychological phenomenon), where the mental facts (concepts) are associated with representations of the linguistic sounds (sound images) that are used to express some thought or idea. The sound images then are arranged with reference to the grammatical rules of the language to express their meaning and to make them understandable for the listener. The circuit step now is in the physiological process, where the brain delivers the sound image to the vocal organs for producing sounds. The sound waves then travel from the speaker’s mouth to the listener’s ears; this is a physical process. The circuit now continues in the listener in reversed order, where from the ear to the brain the physiological transmission of the sound image occurs, and in the brain the psychological association of the image with the corresponding concept takes place. The same process with successive steps occurs when the listener then responds to the original speaker. If all goes well, correct speech perception occurs so that both persons have successfully built communication.

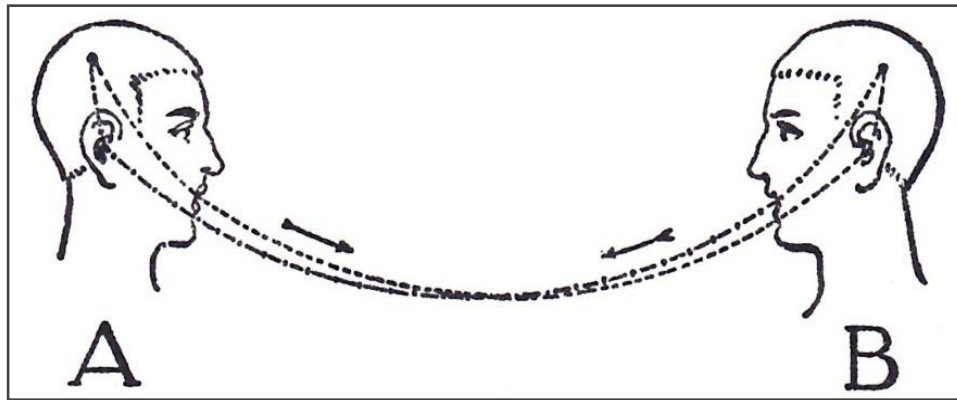


Figure 1.1. The transmission of the linguistic codes from the speaker (A) to the listener (B) (from: de Saussure, 1966).

In the present thesis, speech intelligibility is investigated. Following Figure 1.1, the research starts from the physiological process, where the linguistic sounds are produced. The recorded speech signal was then manipulated. The sound manipulations were then presented to the listener, and information from the sound would be then be converted into a linguistic representation. Before going into detail, further general knowledge about speech and speech intelligibility is described.

1.1 Acoustic Properties of Speech

The acoustic properties of speech sounds can be understood by considering the sound produced from the sound energy sources, such as the vibration of the vocal cords passing through the vocal tract, which modifies the spectrum of the source (Diehl, 2008). By opening the glottis narrowly, the vocal folds vibrate and can produce “voiced sounds”, i.e., vowels (e.g., /a/ and /u/), nasal consonants (e.g., /m/), liquids (e.g., /r/ and /l/) and glides (e.g., /w/). When the glottis opens widely, as in normal breathing, the vibration of the vocal folds is reduced. This can produce “voiceless sounds or aspiration”, i.e., the aspirated /h/, the fricatives /f/ and /s/, and the stop consonants /p/ and /t/. Some speech sounds have multiple sources operating at the same time or sequentially, i.e., the fricative /z/ is produced with a voiced source and a

simultaneous turbulence noise (e.g., frication) source, while the stop consonant /t/ may be produced with a temporary obstruction, a frication and an aspiration source, successively, as the mouth opens (Fant, 1973).

It is well known that speech is a complex acoustic signal that widely varies over time. The signal consists of many different acoustic and linguistic properties that may in different degrees be informative for understanding the intended message. Some acoustic cues may be correlated to provide similar information for understanding speech and probably many cues provide different information too. For understanding speech, those acoustic cues serve different roles (Fogerty & Humes, 2012). Several studies have explored the roles of the acoustic properties of speech (e.g., amplitude envelope, temporal fine structure, and fundamental frequency) in consonant and vowel segments. The speech envelope consists of amplitude modulations that vary relatively slowly over time (Fogerty & Humes, 2012). These temporal amplitude cues facilitate word prediction above what is provided by phonetic information alone (Waibel, 1987). Furthermore, envelope information is important to convey manner and voicing cues (Apoux & Bacon, 2004; Gallun & Souza, 2008; Rosen, 1992; Shannon et al., 1995). The importance of envelope information for general speech intelligibility has been established as well (e.g., Dorman et al., 1997; Shannon et al., 1995). For example, in Shannon et al.'s (1995) study, speech intelligibility was at a fairly high level for word identification in sentences, as well as for consonants and vowels with envelope cues in only three different spectral bands of modulated noise.

Temporal fine structure (TFS) conveys a relatively fast modulation of frequency over time. The TFS information has been found to be most important to recognize the consonant place articulation, in quiet and in noise (Apoux & Bacon, 2004), or in hearing-impaired and normal listeners (Rosen, 1992). Rosen (1992) proposed that the temporal fine structure may best capture the most varied parts of the vowel, i.e., formant transitions, supported by the

evidence that TFS conveys place cues, which provide important information about neighboring consonants (e.g., Liberman et al., 1967). The TFS cues may be able to extract the consonant-information in vowels and lead to probabilistic linguistic cues in sentences (Fogerty & Humes, 2012). Furthermore, some authors argued also that the reception of unknown sentences (Hochmair-Desoyer et al., 1980, 1985) or the accurate identification of articulation place in consonants (Shannon et al., 1992) implies some linguistic use of the TFS of speech.

Another important acoustic property of speech is fundamental frequency (F_0), which is known as an acoustic measure reflecting the rate of vocal fold vibration (Baker et al., 2008). The correlation between F_0 and speech intelligibility has been investigated by several researchers. Laures and Weismer (1999) found that sentence intelligibility was low when the F_0 contour was smoothed, but intelligibility was higher for vowel-only sentences, even though the dynamic cues of the F_0 were removed. Fogerty and Humes (2010) also found that F_0 may be a potential cue that vowels are more apt to carry information than consonants, and would likely facilitate the perception of sentences rather than isolated words.

1.2 Speech Perception and Speech Intelligibility

As described above, speech perception can occur after the linguistic information went successfully to the listeners' brain (Denes & Pinson, 2015). Simply, speech perception can be defined as the process of interpretation of a language sound that is processed by the human auditory system into a linguistic representation (Holt & Lotto, 2010; Smith, 2012). When people are involved in a conversation or communication, they do not simply hear the information conveyed in a sound waveform or a spectrogram, but they perceive linguistic information that conveys words. Then they analyze the words and interpret them. In other words, perception is closely related to how speech is produced or articulated (Casserly & Pisoni, 2010) and how human² hearing and cognitive functions work (Nusbaum & Schwab,

1986; Nusbaum & Magnuson, 1997; Heald & Nusbaum, 2014). Moore et al. (2008) concluded that speech perception depends strongly on the dynamic nature of the sounds of speech and the way that they change over time. In the last decade, many authors have studied speech perception and typically used a single speech sound (e.g., vowels or consonants; Liberman, 1956; Whalen, 1989; Moore et al., 2007; Diehl, 2008), syllables (Nearey, 1990; Greenberg & Arai, 2004), spoken words (Garrett, 1978; Pisoni et al., 1985; Nygaard & Pisoni, 1998; Holt & Lotto, 2010; Jeddi et al., 2012; Moradi et al., 2014), or sentences (Nygaard & Pisoni, 1998; Jeddi et al., 2012) .

Speech intelligibility can be defined as a measure of how understandable speech is in given conditions. In other words, speech intelligibility reflects how clearly a person speaks so that his or her speech can be understood by others (Leddy, 1999). This is important in speech perception; reduced speech intelligibility leads to misunderstanding and loss of interest by communication partners in daily life.

1.3 Characteristics of English Speech and Its Intelligibility

English is an international language used by many people in daily life communication, who either use English as a native language (ENL; e.g., people from Ireland, Australia and Canada; Trudgill & Hannah, 2017), English as a second language (ESL; e.g., people from Singapore, Nigeria and Kenya; Trudgill & Hannah, 2017), or English as a foreign language (EFL). Although English is spoken all over the world, it is sometimes complicated to define English speech sounds, due to phonetic varieties in different geographic and social environments (Wells, 1982). There are many different ways to pronounce sounds in English, depending on accent preferences or personal habits. In some cases, there can be many indistinguishable pronunciations of different consonants, such as between /θ, “th”/ and /f/, and

/d/ and /ð, “th”/. English allows a lot of allophonic variants of each phoneme, and this makes more confusion when the language is pronounced in two or more ways (Carley et al., 2018). Differences in English speech sounds are even larger among non-native speakers of English, who often refer to pronunciation patterns in their own language accent (Volín & Skarnitzl, 2018). For example, Indonesian speakers sometimes pronounce English words in a specific way since their native language, which is written with Roman letters, has a high degree of grapheme-phoneme correspondence, so that words are pronounced as they are written. English, however, is more non-phonemic (Wenanda & Suryani, 2016).

Over the past decades, intelligibility of English speech has been studied with various stimulus patterns. Sentences are more intelligible than individual words (Dirks et al., 1969; Shafiro et al., 2011; Kidd & Humes, 2010; Shafiro et al., 2016), as was shown in gated and temporally-altered English speech. A word will be more intelligible when it is presented in a sentence, because listeners can identify it from the English syntactic structure (Miller & Isard, 1963) and the semantic context in congruent sentences (Kalikow et al., 1977), which can speed up target-word identification in comparison with word-alone presentation (Miller et al., 1951; Grosjean, 1980; Salasoo & Pisoni, 1985). By these findings, because of the importance of English as a tool of communication in today’s world, in the present thesis research is performed on the temporal aspects of English speech processing.

1.4 Speech Intelligibility in Environmental Noise and Reverberation

Most of the research on speech intelligibility so far has been performed with speech under adverse conditions. The intelligibility of speech is adversely affected when the speech is accompanied by other sounds, such as noise, or reverberation. For example, when we are talking in a party room, a restaurant, a subway, an air plane, and so on, important

communication, such as speech announcements, are sometimes difficult to hear. This is a general problem for everyone, but especially for people with central auditory processing disorder, a learning disability, attention deficit, or hyperactivity disorder, for many elderly people over 65 years (Chermak & Musiek, 1997; Bogardus, Yueh & Shekelle, 2003; Warrier et al., 2004; Souza et al., 2007; Anderson et al., 2010), and for non-native listeners of the language in which the communication is made (Broersma & Scharenborg, 2010). The presence of such noise or reverberation in the surrounding area causes difficulty in understanding the speech, regardless of whether the listener is wearing a hearing aid or not (Killion, 1997; Moore, 1998). In this case, the speech information can be conveyed correctly when the speaker increases his/her vocal effort (Summers et al., 1988) or when the listener's hearing and cognitive system (Nusbaum & Schwab, 1986) works strongly to understand the speech. Understanding how noise and reverberation or other sounds influence speech intelligibility is very important.

For people with normal hearing, speech can often be understood even when there is background noise (Yoo et al., 2007; Darwin, 2008; Crespo & Henriks, 2014). Noise can reduce the intelligibility of speech, but has no such effect when the intensities of the speech and the noise exceed a 20-dB signal-to-noise ratio (Denes & Pinson, 2015). However, our experiences in daily life with various speech-under-noise situations show that speech is often intelligible even when its intensity is lower than that of noise (Mei & Sun, 2001; Denes & Pinson, 2015). For example, when we are in a conversation on a busy street, our perceptual mechanism in some way arranges to separate the speech and the noise from the street. In such situations, the listener can also benefit from the "Lombard effect", which means that the speaker modifies his/her speech to make it robust against noise, and also reverberation (Lane & Tranel, 1971).

Besides noise, reverberation quite often accompanies the speech sounds in everyday conversation, and affects the intelligibility performance too. Reverberation as produced by

early and late reflections of the signal, blurs temporal and spectral cues and flattens formant transitions (Nabelek, 1993). As with noise, the speech under reverberation also could be understandable for people with normal hearing (Neuman & Hochberg, 1983; Crespo & Henriks, 2014; Dong & Lee, 2018), but its intelligibility can be reduced. For example, recognition of words deteriorates considerably for hands-free (e.g., telephone) speech input in reverberant environments because the reverberation masks the spectral features of certain phonemes. The vowels typically mask the following phonemes (Hirsch, 1992). Perception of phonemes is degraded in reverberant conditions, e.g., school classroom environments, for children (7-14 years old), but improved for adults (23-30 years old) (Neuman & Hochberg, 1983; Wróblewski, 2012). In other words, children needed better acoustic environments to reach equivalent sentence recognition with their older peers and adults. A similar situation holds for people with hearing impairment (Harris & Swenson (1990).

1.5 Intelligibility of Temporally Manipulated or Segmented Speech

A wealth of research has been performed on temporal aspects of speech processing, by using speech in which parts of the signal were omitted or segmented into units. An early study on the perception of distorted speech employed speech in which 50-ms portions were alternately played and silenced (Miller & Licklider, 1950). Surprisingly, despite the silent gaps, listeners could still extract some meaning from the signals. Further studies on such “gated speech” showed that even if the 50-ms silent gaps were removed and the remaining speech portions were contracted, the speech could still be intelligible (Fairbanks & Kodman, 1957; Shafiro et al., 2016). Besides periodically interrupted speech, processing of distorted speech has further been investigated with speech that was temporally smeared (Drullman, Festen, & Plomp, 1994a,b) or temporally reversed (Kellogg, 1939; Meyer-Eppler, 1950).

Of particular interest is the perception of locally time-reversed speech. In locally time-reversed speech, speech was segmented into short portions of, for example, 50 ms. Following this, each segment was reversed in time, connected again, and presented to the listener (Steffen & Werrani, 1994; Saberi & Perrot, 1999). Studies have shown that the intelligibility of locally time-reversed speech was near zero when the segmented portions were about 100 ms or longer. For shorter segments, however, intelligibility sharply increased and became very high (> 90%) for segments of about 40 ms or shorter, if the speech rates were normalized (Ueda et al., 2017; Nakajima et al., 2018). A study with “pixelated speech” also showed that German speech with a segment duration of 50 ms or shorter obtained almost the same intelligibility as the original speech (Schlittenlacher et al., 2019).

1.6 Neuroscientific Findings Related to Temporal Aspects of Speech Processing

During the last decades, auditory neuroscience research has added new insights into temporal aspects of speech processing, by proceeding from speech units based on phonetic segmentations (Liberman & Mattingly, 1985), articulatory features (Stevens, 2002), or syllables (Nearey, 1990; Greenberg & Arai, 2004). Especially the importance of neural oscillations in cortical speech processing has been stressed, in particular of those with a modulation frequency-range around 30-50 Hz (Giraud & Poeppel, 2012). Neural oscillations with this modulation frequency are thought to be engaged in phonemic processing (Ding et al., 2017). Interestingly, the modulation frequency of these neural oscillations corresponds to a temporal window of around 20–33 ms (Chait et al., 2015), which corroborates the idea that the human auditory system processes speech in relatively rough time segments. Both neuroscientific studies and studies based on psychophysical methods on locally time-reversed speech thus suggest that the duration of these time segments is about 40 ms or smaller.

1.7 Mosaic Speech

Recently, Nakajima et al. (2018) introduced a new type of speech stimulus, called “mosaic speech”. It has been developed to further study speech processing in general and its temporal acuity in particular. This mosaic speech adopted the basic principle of making a mosaic image of a visual image (Harmon, 1973). Since a visual image can be mosaicked and still can be shown and seen (Figure 1.2a), Nakajima et al. (2018) thought that speech can be mosaicked as well to be played and heard (Figure 1.2b).

Investigating speech in mosaicking offers a possibility to measure the temporal resolution that humans need for speech perception. One purpose of using “mosaic speech” was to provide an alternative to locally time-reversed speech (Steffen & Werrani, 1994; Saberi & Perrot, 1999; Ueda et al., 2017; Nakajima et al., 2018). Local time-reversal can leave some unintended cues, as well as distortions, about the spectral content of the speech signal. By performing listening experiments with mosaicked Japanese speech, in which the frequency resolution was as fine as a critical bandwidth, Nakajima et al. (2018) found that intelligibility was near-perfect (> 95%) at shorter block durations up to 40 ms. This finding was similar to the results of Kojima and Nakajima (2016) and Kojima et al. (2017), in which the perception of English mosaic speech (sentences) was investigated, with Japanese listeners. At a segment duration of 40 ms, about 80% of the speech was understood. Intelligibility of mosaic speech decreased dramatically at longer block durations, from 40 to 320 ms (Kojima & Nakajima, 2016; Kojima et al., 2017; Nakajima et al., 2018).

Studies with mosaic speech and locally time-reversed speech (Ueda et al., 2017; Nakajima et al., 2018) thus showed the highest intelligibility at segment durations of 40 ms or less. Although the intelligibility of locally time-reversed speech and the intelligibility of mosaic speech were similarly dependent on segment duration, mosaic speech, for which intelligibility was systematically higher, is considered as more suitable than locally time-reversed speech to

investigate the temporal nature of speech perception. In locally time-reversed speech, the content of each reversed temporal segment is never static. However, in mosaic speech, the blocks are static (except for the random fluctuation of noise), and have a frequency resolution as fine as a critical band.

In brief, mosaic speech was made in the following way (for further details about the generation of mosaic speech, the reader can refer to section 2.2.3). The initial procedure resembled the procedure to generate noise-vocoded speech (Shannon et al., 1995; Smith, Delgutte, & Oxenham, 2002; Ellermeier et al., 2015; Kishida et., 2016). First, the original speech signal was separated into several frequency band-pass filters, mimicking the auditory periphery, which is considered to work as if made of non-overlapping, but closely packed frequency bands called critical bands. The waveform of each frequency band was cut into temporal segments of the original mosaic block duration (OMBD), for example of 40 ms, the total amount of its sound energy was calculated by squaring and adding up instantaneous amplitudes. A white noise of the same duration as that of the speech signal was generated and went through the same frequency band-pass filters. The waveforms in these band-pass filters then were cut into temporal segments. Cosine-shaped rise and fall times were used around each temporal segment or block, and the total amount of its sound energy was calculated in the same way. There was a time-frequency correspondence between the speech signal and the white noise, because they had the same duration and the same frequency range. Each temporally segmented band noise was amplified to make its total sound energy equal to that of its counterpart in the processed speech signal. By putting all these amplitude-adjusted segmented band noises together on the time-frequency plane, mosaic speech was obtained (Figure 1.2).

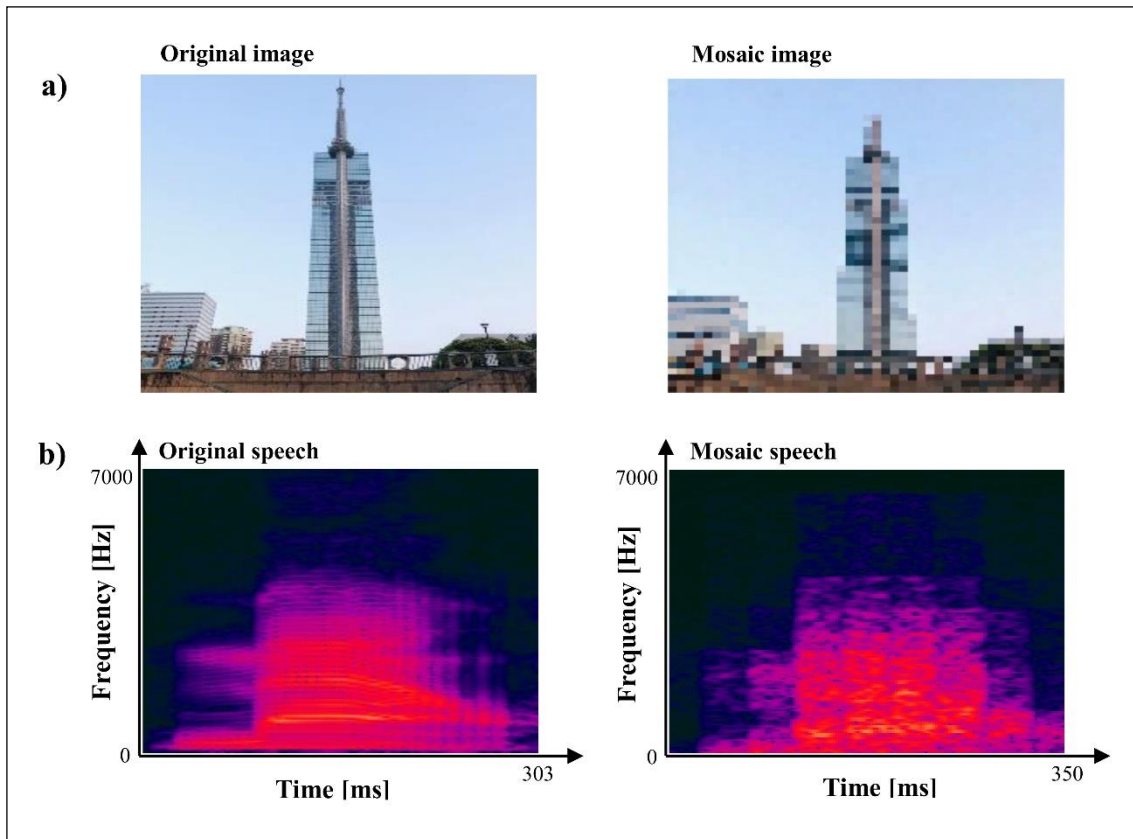


Figure 1.2. An original visual image (a, left) changed into mosaic image (a, right) of Fukuoka Tower, Fukuoka, Japan. A mosaic image can be created by averaging the color values or luminance grades within each block. An original speech (b, left) was changed into mosaic speech (b, right) by averaging the total amount of sound energy (as indicated by yellow and orange colors; taken from the Cool Edit 2000 software window).

1.8 Compressed and Stretched Speech

Temporal information processing plays a fundamental role in speech processing (Hickok & Poeppel, 2007; Szelag et al., 2004a). Experimental findings support the concept that neural representations of single units of language, like phonemes, are related to specific mechanisms that also control temporal processing. Although individual differences in speed of speech may vary considerably, an extremely fast or a very slow speed can have a disturbing effect on speech understanding. In daily life communication, for example in my case as an Indonesian listener, listeners are quite often faced with situations in which it is difficult to understand what other people are saying, especially when they are speaking quickly using

another language, like English. It is common that the speaking speed affects the intelligibility of the speech. The utterance speed of English speech should be below 400 words per minute for effective communication (Du, Lin & Wang, 2015). When the verbal output becomes too fast or too slow, the listener cannot process speech utterances easily and will not be able to extract meaning adequately. Meanwhile, for speech uttered in normal speed, a study with the gating paradigm showed that word identification in general occurs after a little more than half of the total word duration has been presented (Grosjean, 1980; Salasoo & Pisoni, 1985).

From several years ago, the relationship between the speed of speech and speech intelligibility has been investigated by modifying the time-scale of that speech. When compressing or stretching English speech, even people with normal hearing prefer speech rates that have been slightly slowed in comparison to average conversational speech rates, which typically are in between 140 and 180 words per minute (Wingfield, 1996; Wingfield & Ducharme, 1999; Wingfield et al., 1999; Wingfield et al., 2006). This preference for and improved performance with slightly slowed speech rates has been particularly documented in degraded listening conditions (Beasley et al., 1972; Konkle et al., 1977; Schmitt & McCroskey, 1981; Schmitt, 1983; Wingfield & Ducharme, 1999; Moore et al., 2007). A similar trend has been shown for listeners with hearing impairment, showing that performance on speech understanding measures decreases with an increase in the speed of speech (Luterman et al., 1966; Sticht & Gray, 1969; Gordon-Salant & Fitzgibbons, 2001; Versfeld & Dreschler, 2002).

1.9 General Purpose

In the present thesis research, several aspects of the perception of mosaic speech are being investigated in order to better understand human processing of speech in general. First, it has been shown that mosaic speech with an OMBD of 40 ms or shorter was almost perfectly

intelligible (> 95%) for Japanese speech (Nakajima et al., 2018), and intelligibility was around 80% for English speech presented to foreign speakers (Kojima & Nakajima's, 2016; Kojima et al.'s, 2017; see Section 1.7). The key question is: Does it matter how much linguistic information is carried in segments of 40 or 20 ms? One way to investigate this is to compress or stretch the same linguistic information in time. Mosaic speech is very suitable for this. As mentioned in Section 1.5, there are many ways to measure the temporal resolution of speech that is necessary to preserve sufficient information for auditory perception, i.e., gated speech (Fairbanks & Kodman, 1957; Shafiro et al., 2016), temporally-smearred speech (Drullman, Festen, & Plomp, 1994a,b), or locally time-reversed speech (Steffen & Werrani, 1994; Saberi & Perrot, 1999, Ueda et al., 2017; see also Kellogg, 1939; Meyer-Eppler, 1950). As discussed in Section 1.7, in mosaic speech the blocks are static, and thus are suitable to be compressed or stretched.

So far it is not known whether intelligibility would be similarly good when the OMBDs are compressed or stretched in time (i.e., when the mosaic blocks are made shorter or longer), while preserving the same acoustic information. For example, if original mosaic blocks of 80 ms, which are not sufficiently understandable, are compressed into mosaic blocks of 20 ms, does this improve intelligibility? As an opposing example, if original mosaic blocks of 20 ms, which are reasonably intelligible, are stretched out into mosaic blocks of 80 ms, does intelligibility deteriorate? Thus, in order to achieve the general purpose, the first objective of the present thesis research was to measure the intelligibility of mosaic speech in which OMBD was either compressed, preserved, or stretched in time.

The secondary objective of this present thesis research was to investigate whether the intelligibility of mosaic speech varies with the language background of the listener. Instead of Japanese, as used previously (Nakajima et al., 2018), English mosaic speech was used here. Whereas near-perfect (> 95%) intelligibility of mosaic speech with segments of 40 ms or

shorter has been reported for the Japanese language (Nakajima et al., 2018), the present thesis research investigated whether English mosaic speech, with all its complexity in speech sounds as described above (see Section 1.3), would also be near perfect at a similar segment duration or not. Apart from preliminary studies (Kojima & Nakajima, 2016; Kojima et al., 2017), no systematic data on the perception of English mosaic speech have been gathered. Therefore, the intelligibility of English mosaic speech was investigated for listeners with three different language backgrounds, i.e., Indonesian (in the preliminary and in the main experiment), native-English, and Chinese (in the main experiment).

1.10 Structure of the Dissertation

Chapter 2 describes a preliminary experiment that investigates which block durations of English mosaic speech were able to convey intelligible linguistic information. In this chapter, all research methods that are used have been described, e.g., how the speech stimuli were selected, how the speech stimuli were recorded, how the speech sounds were mosaiced, and last, how the stimuli were presented to the listeners. In the preliminary experiment, the intelligibility of English mosaic speech was investigated with OMBDs of 20 and 40 ms after compression, preservation, and stretching. Twenty Indonesian participants were employed and the analysis of the qualitative data was conducted through non-parametric tests, e.g., a Friedman two-way analysis of variance by ranks, a Wilcoxon signed-rank test and Holm-Bonferroni tests as post-hoc tests (Field, 2009).

Chapter 3 describes the main experiment that further investigates the preliminary experiment's finding. The experiment consists of 3 sub-experiments, one for each of three language groups. The experiment was performed with a native-English group of listeners, an Indonesian group, and a Chinese group. In these experiments, almost the same procedures as in the preliminary experiment were used, and the analysis methods were the same as well.

Besides describing the intelligibility of mosaic English words, this chapter describes also the intelligibility of the phoneme of the initial consonant of each word. The General Discussion and Conclusions are described in Chapter 4. This chapter summarizes the findings and provides a discussion based on the results of Chapters 2 and 3, and it also describes areas for further studies.

Chapter 2 - Preliminary Experiment: Effects of Compressing or Stretching Mosaic Block Duration on English Speech Intelligibility

2.1 Purpose

The previous studies on mosaic speech made with English sentences (Kojima & Nakajima, 2016; Kojima et al., 2017) and Japanese sentences (Kojima et al., 2017; Nakajima et al., 2018) showed that the intelligibility of mosaic speech was highest with an OMBD of 20 and 40 ms. As a preliminary investigation, these block durations were used to investigate the intelligibility of English words, after compressing, preserving, or stretching those OMBDs. The first purpose of this experiment was to investigate which mosaic block durations would be appropriate for English speech stimuli for the main experiment. Non-native speakers of English were employed. The second purpose was to examine whether all experimental procedures were in order, e.g., with regard to stimulus recording and stimulus presentation.

2.2 Method

2.2.1 Participants

Twenty non-native English speakers (Indonesian speakers) were employed as listeners. They were 8 men and 12 women (20-38 years old). All were students of Kyushu University with normal hearing. The participants' scores on English proficiency tests were as follows. Five participants had taken the International English Language Testing System (IELTS; scores = 6.5-8.0). Fifteen participants had taken the Test of English as a Foreign Language (TOEFL;

scores = 503-577). All participants agreed to participate and provided written informed consent. The experiment was conducted with prior approval of the Ethics Committee of Kyushu University.

2.2.2 Equipment

Stimulus recording

Stimuli (English words) were recorded from a native-English speaker in a soundproof room (of which the background noise level was about 25 dBA), by using a digital recorder (TASCAM, DR-07, Teac Corporation, Tokyo, Japan), covered by a pop filter and placed on a tripod. A sound level meter (ACO, Type 6240, ACO Co, Ltd., Tokyo, Japan) was used to monitor the sound level of the spoken words.

Stimulus presentation

The experiment was conducted in the same soundproof room as used for recording. The stimuli were stored in a computer (ONKYO, M513A8, ONKYO Corporation, Tokyo Japan) that was placed outside the room. From the computer, the stimuli were passed through an audio interface (Roland, UA-1010), a low-pass filter (NF DV-04 DV8FL, NF Corporation, Yokohama, Japan; cut-off frequency 15 kHz), a graphic equalizer (Roland, RDQ-2031), and a headphone amplifier (STAX, SRM-3235, STAX Limited, Saitama, Japan), before diotical presentation through headphones (STAX, SR-307). The presentation level of all stimuli ranged in between 66–75 dBA (Fast-Peak), as measured by using an artificial ear (Brüel & Kjør, 4153, Nærum, Denmark) and a sound level meter (ACO, Type 6240).

2.2.3 Stimuli

English word specifications

Eighty English words in a structure of Consonant-Vowel-Consonant (CVC), Consonant-Vowel (CV) or Vowel-Consonant (VC) were used. The words were derived from English textbooks for children up to elementary school (Kampa & Vilina, 2001; Rivers & Toyama, 2011) within the category of “content words”, which have lexical meaning (Richards & Schmidt, 2010). Some criteria were applied to avoid any ambiguity in word meaning. The words were selected as follows. Words ending with the letter “r” were not considered, since its pronunciation can sometimes be lost, which is known as a non-rhotic accent (e.g., four; [fɔ:(r)]; (Carley et al., 2018). Furthermore, words with two or more possible pronunciations depending on dialect were excluded (e.g., dog, [/dɒg/] or [/dɔg/]; Wells, 2008). Finally, the 100 most commonly used words (Kress & Fry, 2016) were excluded as well, since the words can have several meanings depending on the function word accompanying them (e.g., “look”; “look at”, “look for”, “look down on”, and “look forward to”). All selected words were included in the 3000 most frequently used words in both spoken and written English (Wells, 2014). The selected words are shown in Table 2.1.

Table 2.1. The 80 words used in the English mosaic speech preliminary experiment.

Group	Word	Group	Word	Group	Word	Group	Words
A	bag	F	hide	K	keep	P	fight
	case		lack		nine		name
	fill		page		raise		pan
	red		sit		win		sheep
B	leaf	G	bus	L	feed	Q	book
	moon		cat		gate		ten
	rate		lake		love		late
	young		wide		map		sad
C	date	H	hit	M	bed	R	five
	house		leg		cut		hat
	king		pain		pick		pull
	match		touch		shape		tape
D	hate	I	fan	N	fat	S	nice
	kick		night		june		cook
	boy		push		egg		fun
	ride		shake		sing		wave
E	big	J	eight	O	catch	T	fish
	cake		lip		put		line
	heat		mad		rain		pay
	wine		run		seed		seat

Stimulus recording

All words were pronounced by a male, native-English speaker (from the United Kingdom, age = 28 years old). The word list was divided into twenty groups, each containing 5 words. The list was stuck to the wall in a soundproof room. The speaker sat on a chair and the recorder was placed in front of the speaker and was fixed on a tripod with a soft material between them. Other soft materials also were placed between the tripod and the floor. In order to avoid sudden sound bursts reaching the recorder, a pop filter was placed about 5 cm from the speaker's mouth. A sound level meter was put below the recorder. The speaker practiced to read the list first to avoid mistakes during the recording. The list was read per group, and the speaker read it word by word (3 times for each word) with a silence of about 2-5 seconds between them. The reading was started about 2 seconds after the record button was pressed for each group. A silence of about 30 seconds was placed as a space between groups (Appendices

A and B). Speech sounds recordings were stored as a library of wav files using a sampling rate (SR) of 44100 Hz, 16 bit, and mono-channel. One sound of each word was selected as a stimulus to be used in the experiment. The selection was based on having limited fluctuation in the speech signal amplitude (within -3 or $+3$ dB overall) and the phonemes were checked by use of the Cambridge Dictionary online (<https://dictionary.cambridge.org/>, accessed on 26 July 2019). Furthermore, an empty duration of 10 ms was added before each word and 5 ms after the end of each word.

English speech mosaicization

Mosaic speech generation could not proceed from any existing mosaicization program as for visual images (Harmon, 1973). In sound, an uncertainty principle works between time and frequency, the inverse of time. Therefore, it is impossible to cut both the horizontal and the vertical axis of the sound spectrogram into pieces very accurately. To cope with this, mosaic speech was generated by constructing an algorithm in which temporal resolution of 20 ms was secured, considering the fact that the period of vocal-folds vibration of male speakers can be around 10 ms (Raphael et al., 2011). Frequency resolution of 50 Hz, higher than the narrowest critical band (Fastl & Zwicker, 2007), was secured as well (see Nakajima et al., 2018). The recorded words were thus transformed into mosaic speech stimuli with an in-house made program written in the “J” programming language.

Using this computer program, the original speech signal was separated first into 20 band-pass filters, mimicking the auditory periphery, which is considered to work as if made of non-overlapping, but closely packed frequency bands called critical bands, covering a frequency range of 50–7000 Hz (Fastl & Zwicker, 2007) (Figure 2.1a,b). All waveforms in these band-pass filters were cut into temporal segments of the original mosaic block duration (OMBD). An example of 40-ms segments is shown in Figure 2.1c. As mentioned earlier, in the

previous study (Nakajima et al., 2018), it was found that the intelligibility of Japanese mosaic speech was near-perfect ($> 95\%$) if the mosaic block duration was 20 or 40 ms. Therefore, the OMBDs of 20 and 40 ms were selected for the present thesis research as well.

As the next step in generating mosaic speech, the total amount of sound energy in each temporal segment in each frequency band was calculated by squaring and adding up instantaneous amplitudes. A white noise of the same duration as that of the speech signal was then generated. It went through the same band-pass filters, and the waveforms in these band-pass filters were cut into temporal segments in the same way. For each temporal segment, cosine-shaped rise and fall times of 4 ms were used. The total amount of sound energy in each temporal segment in each frequency band was calculated in the same way. There was a time-frequency correspondence between the speech signal and the white noise, because they had the same duration and the same frequency range. As a following step, each temporally-segmented band noise was amplified, disamplified, or kept unchanged to make its total sound energy equal to that of its counterpart in the speech signal as processed above. Mosaic speech was obtained by putting all these amplitude-adjusted segmented band noises together on the time-frequency plane.

For the present experiment, the OMBD was compressed (Figure 2.1d), preserved, or stretched (Figure 2.1e) by reducing, keeping, or increasing the number of samples for each mosaic block. The OMBD was compressed into half ($0.5 \times \text{OMBD}$), preserved ($1 \times \text{OMBD}$), or stretched by a factor of 2, 4, or 8 ($2 \times \text{OMBD}$, $4 \times \text{OMBD}$, $8 \times \text{OMBD}$). The resulting duration was called “Mosaic Block Duration” (MBD); the shortest MBD was 10 ms ($0.5 \times \text{OMBD}$ of 20 ms), and the longest was 320 ms ($8 \times \text{OMBD}$ of 40 ms), as indicated in Table 2. The spectral pattern and the power level inside the MBD after compressing/preserving/stretching remained the same, ensuring that each block contained the same acoustic information.

Table 2.2. Mosaic speech block durations used in the preliminary experiment

MBD after compressing/preserving/stretching	OMBD: 20 ms		OMBD: 40 ms	
	Mosaicizing phase type			
	Half (10 ms)	Whole (20 ms)	Half (20 ms)	Whole (40 ms)
Compressed (OMBD \times 0.5)	10	10	20	20
Preserved (OMBD \times 1)	20	20	40	40
Stretched 2 (OMBD \times 2)	40	40	80	80
Stretched 4 (OMBD \times 4)	80	80	160	160
Stretched 8 (OMBD \times 8)	160	160	320	320
	(ms)	(ms)	(ms)	(ms)

There were two mosaicizing phase types, the half-phase type and the whole-phase type. Since an empty duration of 10 ms was already added at the beginning of the original speech signal, another portion of empty duration was added to make it a half or a whole length of the OMBD, as indicated in Table 2.2 (“Half” and “Whole”). By using two different lengths of the total added duration (10 and 20 ms for the OMBD of 20 ms; 20 and 40 ms for the OMBD of 40 ms), it is possible to explore whether phoneme perception would be affected if the mosaicization began a half-block duration or one block duration earlier than the onset of the speech.

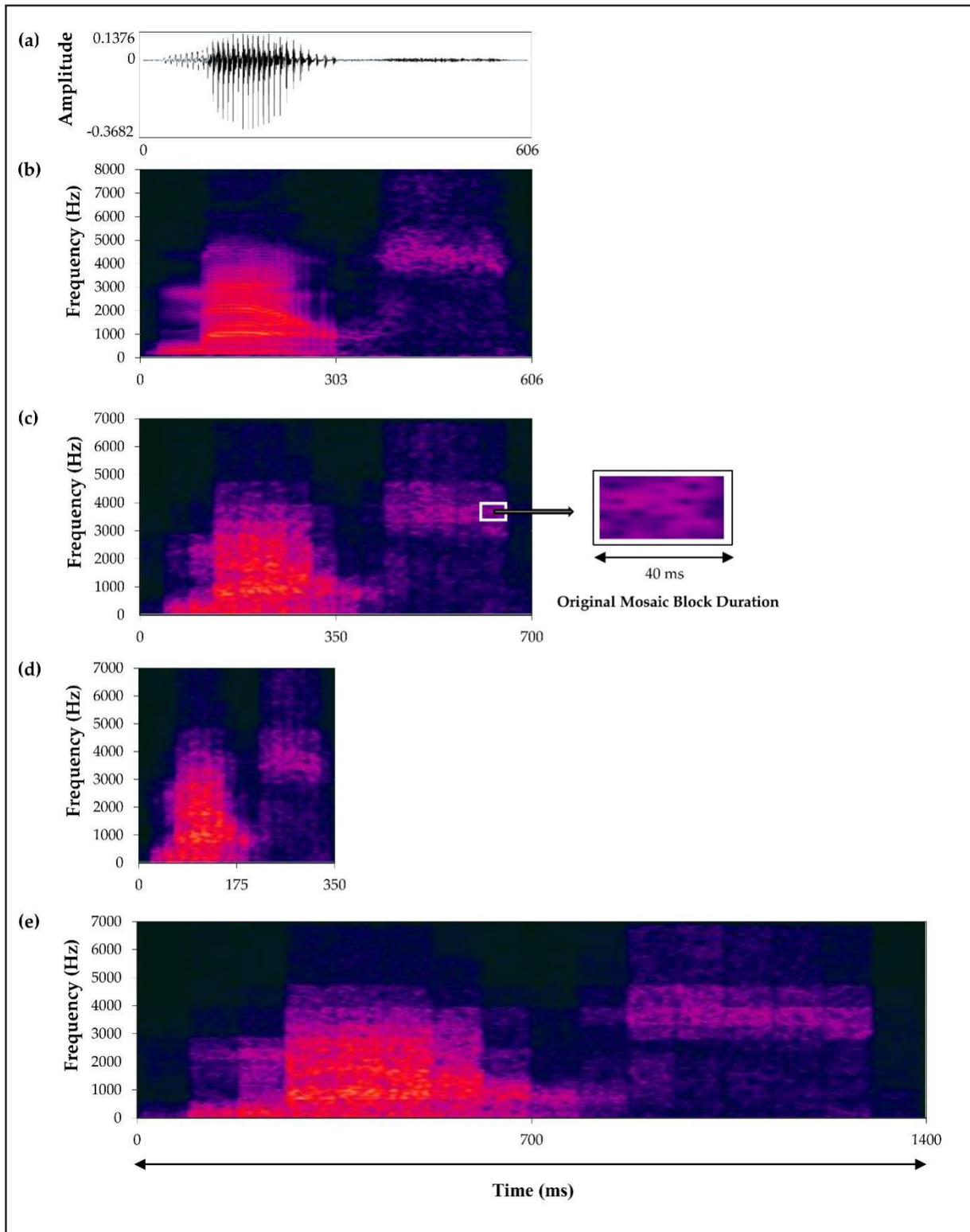


Figure 2.1. Examples of the mosaic speech stimuli used throughout this thesis. (a) The waveform and (b) the spectrogram of the original speech for the word “mouse”, pronounced by a male native-English speaker. (c) An example of mosaic speech with an original mosaic block duration (OMBD) of 40 ms. Each individual block consisted of one mosaic block duration on the horizontal axis and one frequency band on the vertical axis. (d) An example of compressed mosaic speech with an OMBD of 40 ms compressed into a mosaic block duration (MBD) of 20 ms. (e) An example of stretched mosaic speech consisting of an OMBD of 40 ms stretched into an MBD of 80 ms.

2.2.4 Procedures

The 80 words were divided into twenty groups, each containing four words (Table 2.1). Each group was assigned to a different mosaic speech stimulus type, and this assignment was different among participants (see Table 2.3). All participants received all the words, but in different stimulus types. The 20 words that were used for practice trials were also assigned to a different stimulus type, and the assignment was the same among all participants. The intelligibility experiment was started with one block of the practice trials, followed by four main blocks, each containing twenty measurement trials.

The stimuli were presented through headphones in random order to the participant, who sat on a chair in front of the computer interface, which was created in Visual Basic .NET programming language (Visual Studio 2017 version 15.0). The participant was asked to click a “play” button on the interface to start a trial. The stimulus of each trial was presented 0.5 s after the button was clicked. The presentation was repeated three times with 1.5-s intervals. After listening to the sound stimulus, the participant typed the perceived word, if any, using the English alphabet, within 5 seconds. The participant was instructed to avoid guessing the correct answer.

Table 2.3. The assignment of stimulus type to the word groups for each participant, as used in the preliminary experiment.

	Stimulus type																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Participant 1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Participant 2	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	A
Participant 3	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	A	B
Participant 4	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	A	B	C
Participant 5	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	A	B	C	D
Participant 6	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	A	B	C	D	E
Participant 7	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	A	B	C	D	E	F
Participant 8	H	I	J	K	L	M	N	O	P	Q	R	S	T	A	B	C	D	E	F	G
Participant 9	I	J	K	L	M	N	O	P	Q	R	S	T	A	B	C	D	E	F	G	H
Participant 10	J	K	L	M	N	O	P	Q	R	S	T	A	B	C	D	E	F	G	H	I
Participant 11	K	L	M	N	O	P	Q	R	S	T	A	B	C	D	E	F	G	H	I	J
Participant 12	L	M	N	O	P	Q	R	S	T	A	B	C	D	E	F	G	H	I	J	K
Participant 13	M	N	O	P	Q	R	S	T	A	B	C	D	E	F	G	H	I	J	K	L
Participant 14	N	O	P	Q	R	S	T	A	B	C	D	E	F	G	H	I	J	K	L	M
Participant 15	O	P	Q	R	S	T	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Participant 16	P	Q	R	S	T	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Participant 17	Q	R	S	T	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Participant 18	R	S	T	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Participant 19	S	T	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Participant 20	T	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S

2.2.5 Statistical Analysis

Since a Shapiro-Wilk test (Field, 2009) showed that the intelligibility data for all stimulus types were not normally distributed (Appendix D), non-parametric tests were performed for statistical analysis. The Friedman two-way analysis of variance by ranks (Field, 2009) was performed first to analyze the main effect of compressing or stretching the OMBD on the intelligibility scores. Although this was not the main purpose of the present thesis research, the Wilcoxon signed-rank test (Field, 2009) was performed to check whether the use of the two mosaicizing phase types, the half-phase, and the whole-phase, in any of the stimulus types affected the intelligibility. Since no effect of the phase types on phoneme perception of

English words was found and the size effect of the differences between them was small ($r < 0.5$, Cohen 1988, 1992, Table 2.4), for convenience the scores were collapsed.

The Wilcoxon signed-rank test (Field, 2009) was used to further analyze the effect of compressing, preserving, or stretching the 20-ms or 40-ms OMBD within each OMBD by making multiple comparisons. Post-hoc Holm-Bonferroni correction (Holm, 1979; Abdi, 2010) was performed to correct for the number of comparisons between stimulus types and to control the family-wise error rates. The stimuli with an MBD of 320 ms were left out of the analysis, since intelligibility for these stimuli was close to zero. Testing was done with the same statistical methods to compare the intelligibility of the same MBD durations within each OMBD. For example, comparisons were made between the intelligibility of a compressed 40-ms OMBD and a preserved 20-ms OMBD, which both have an MBD of 20 ms. In order to check the effect of compression on intelligibility, for the OMBD of 20 ms we also performed pair-wise comparisons including the compressed condition (MBD = 10 ms). IBM SPSS Statistics (Version 25.0) was used for all statistical computations.

Table 2.4. Results of the preliminary experiment. Results of the Wilcoxon Signed-rank test for comparing the intelligibility between mosaicked stimuli with a half- and a whole-phase type of the OMBD of 20 and 40 ms.

Mosaicizing phase type		* <i>Mdn</i>	Asymp. Sig. (2-tailed)	Sig. ($\alpha= 0.05$)	<i>z</i>	** <i>r</i>
OMBD of 20 ms:						
×0.5 OMBD	Half-phase	50	0.239	Non-significant	-1.18	-0.26
	Whole-phase	50				
×1 OMBD	Half-phase	75	0.506	Non-significant	-0.67	-0.15
	Whole-phase	62.5				
×2 OMBD	Half-phase	50	0.183	Non-significant	-1.33	-0.30
	Whole-phase	75				
×4 OMBD	Half-phase	25	0.287	Non-significant	-1.06	-0.24
	Whole-phase	25				
×8 OMBD	Half-phase	0	0.206	Non-significant	-1.27	-0.28
	Whole-phase	12.5				
OMBD of 40 ms:						
×0.5 OMBD	Half-phase	50	0.414	Non-significant	-0.82	-0.18
	Whole-phase	50				
×1 OMBD	Half-phase	62.5	0.723	Non-significant	-0.35	-0.08
	Whole-phase	62.5				
×2 OMBD	Half-phase	50	0.685	Non-significant	-0.41	-0.09
	Whole-phase	50				
×4 OMBD	Half-phase	0	0.313	Non-significant	-1.01	-0.23
	Whole-phase	25				
×8 OMBD	Half-phase	0	0.564	Non-significant	-0.58	-0.13
	Whole-phase	0				

* *Mdn* is the median of intelligibility of half-phase or whole-phase type stimuli (scale= 0—100).

** *r* is the effect size of the relationship between half-phase and whole-phase stimuli ($r= z/\sqrt{n}$); Rosenthal, 1991).

2.3 Results

Mosaic speech intelligibility was obtained by counting the number of correct answers given by the participants for the Consonant-Vowel-Consonant (CVC), the Consonant-Vowel (CV), and the Vowel-Consonant (VC) words. The word spellings of the participants' answers were automatically matched with the spelling of the actual word stimuli by the computer program used in the stimulus presentation. Figure 2.2 shows the intelligibility scores for the English mosaic speech stimuli. As described above, there was no significant effect of half-phase or whole-phase starting phases on word intelligibility ($r < 0.5$, Cohen 1988, 1992, Table 2.4). Therefore, since this was not the main purpose of the present thesis research, for convenience the scores were collapsed.

Figure 2.2 shows that mosaicing the original speech affected the intelligibility. The intelligibility decreased by about 20% when the OMBD was compressed compared to the intelligibility of the preserved OMBD. The decrease in intelligibility exceeded 50% when the OMBD was stretched by a factor of 4 or longer. Furthermore, the intelligibility did not decrease very much when the OMBD was stretched by a factor of 2, compared to the intelligibility of the preserved OMBD. The intelligibility decreased by only about 3% for the OMBD of 20 ms and by about 18% for the OMBD of 40 ms when the OMBD was stretched by a factor of 2.

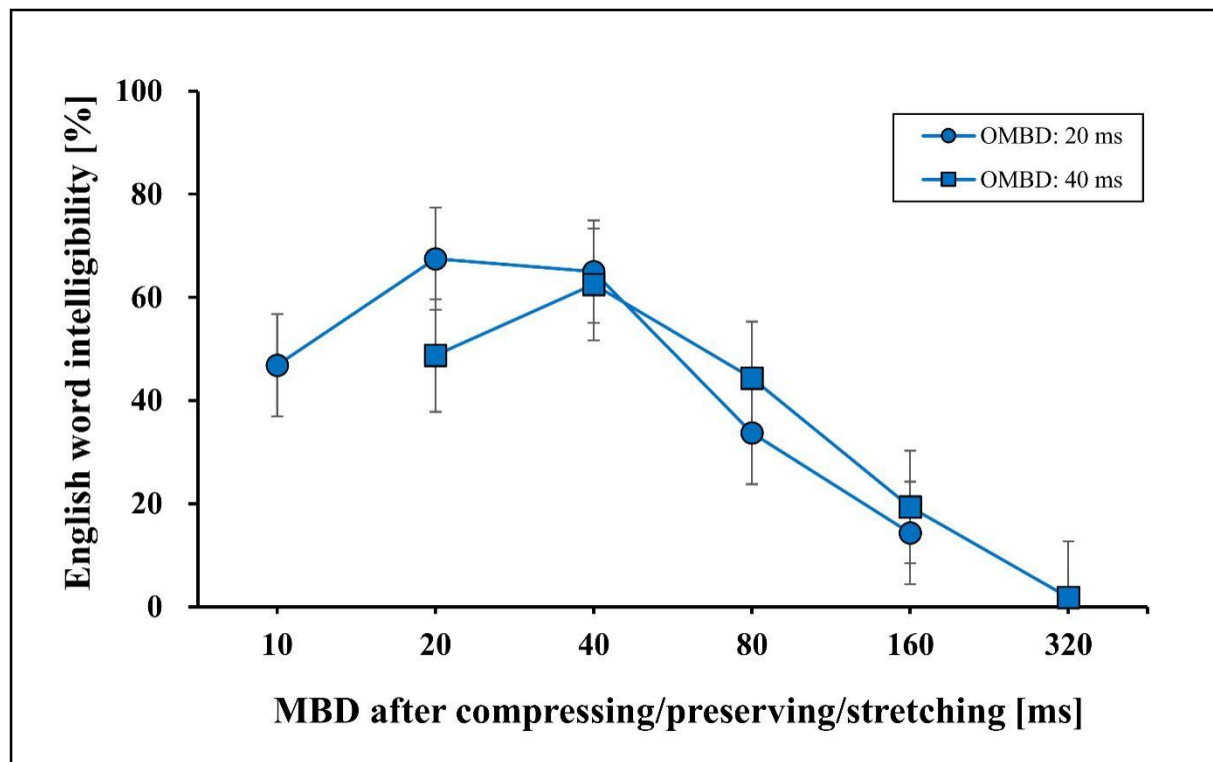


Figure 2.2. Results of the preliminary experiment. English word identification accuracy (intelligibility) for mosaic speech as a function of MBD after compressing/preserving/stretching (Indonesian, $n = 20$). The data for half-phase and whole-phase types are collapsed. Error bars indicate standard errors.

2.3.1 The Effect of Compressing or Stretching the Original Mosaic Block Duration (OMB D)

The intelligibility of English mosaic speech was highest when the OMBDs were preserved, for both OMBDs of 20 and 40 ms. Intelligibility decreased as the MBD was shorter

or longer. When the OMBDs were compressed, the intelligibility decreased by about 20% as compared to the intelligibility for the preserved OMBDs. When the OMBDs were stretched into stimuli with an MBD of 80 ms, the intelligibility decreased sharply, especially for the stimuli with an OMBD of 20 ms. For stimuli with an MBD of 320 ms after stretching, the intelligibility was close to zero.

A Friedman two-way analysis of variance by ranks (Field, 2009) was performed to analyze the main effects of compressing or stretching the OMBD. There was a significant difference in intelligibility between mosaic speech types ($n = 20$, $k = 19$, $\chi^2 = 203.839$; $p < 0.001$). Multiple comparisons with the Wilcoxon Signed-rank test (Field, 2009) were used to follow up on this finding, and a Holm-Bonferroni correction (Holm, 1979; Abdi, 2010) was performed after that. In detail (see Table 2.5), compressing the OMBD degraded intelligibility for the OMBD of 20 ms ($p = 0.004$), but not for the OMBD of 40 ms ($p = 0.067$). The intelligibility of the MBD after preserving was significantly higher than that of the MBD after stretching (80, 160, or 320 ms; $p < 0.05$) for stimuli with both a 20-ms and a 40-ms OMBD. Stretching the OMBD by a factor of 2 led to significantly higher intelligibility than compressing the OMBD by half for the OMBD of 20 ms ($p = 0.012$), but not for the OMBD of 40 ms ($p = 0.336$). Moreover, stretching the OMBDs by a factor of 2 induced higher intelligibility than stretching the OMBDs by a factor of 4 or 8 ($p < 0.01$).

Table 2.5. Results of the preliminary experiment. Intelligibility comparisons between mosaic speech stimuli with a different MBD after compressing, preserving or stretching.

Family of tests (<i>C</i>)		Asymp. Sig. (2-tailed) (* <i>p</i>)	Holm-Bonferroni (<i>C</i> - ** <i>i</i> + 1) × <i>p</i>	Holm-Bonferroni Sig. (α= 0.05)
OMBD of 20 ms:				
×1 vs. ×8	20 and 160-ms MBD	0.000	0.001	Significant (×1 > ×8)
×2 vs. ×8	40 and 160-ms MBD	0.000	0.001	Significant (×2 > ×8)
×1 vs. ×4	20 and 80-ms MBD	0.000	0.001	Significant (×1 > ×4)
×0.5 vs. ×8	10 and 160-ms MBD	0.000	0.001	Significant (×0.5 > ×8)
×2 vs. ×4	40 and 80-ms MBD	0.001	0.004	Significant (×2 > ×4)
×0.5 vs. ×1	10 and 20-ms MBD	0.001	0.004	Significant (×0.5 < ×1)
×4 vs. ×8	80 and 160-ms MBD	0.001	0.006	Significant (×4 > ×8)
×0.5 vs. ×2	10 and 40-ms MBD	0.004	0.012	Significant (×0.5 < ×2)
×0.5 vs. ×4	10 and 80-ms MBD	0.016	0.032	Significant (×0.5 > ×4)
×1 vs. ×2	20 and 40-ms MBD	0.547	0.547	Non-significant
OMBD of 20 ms:				
×1 vs. ×4	40 and 160-ms MBD	0.000	0.001	Significant (×1 > ×4)
×0.5 vs. ×4	20 and 160-ms MBD	0.001	0.003	Significant (×0.5 > ×4)
×2 vs. ×4	80 and 160-ms MBD	0.001	0.003	Significant (×2 > ×4)
×1 vs. ×2	40 and 80-ms MBD	0.012	0.036	Significant (×1 > ×2)
×0.5 vs. ×1	20 and 40-ms MBD	0.033	0.067	Non-significant
×0.5 vs. ×2	20 and 80-ms MBD	0.336	0.336	Non-significant

**p* = *p*-value from multiple comparisons with the Wilcoxon Sign-rank test.

***i* = rank of the family of tests (the smallest *p*-value is the largest rank number).

2.3.2 Intelligibility Comparisons between Stimuli with the Same MBDs

By stretching or compressing, the mosaic speech stimuli with a 20-ms and a 40-ms OMBD contained blocks of the same duration, which are 20, 40, 80, or 160 ms. Given this, it is necessary to determine whether the word intelligibility for stimuli with the same MBD would be similar or not. In detail (see Table 2.6), regarding the stimuli with an MBD of 20 ms, the results showed that compressing the 40-ms OMBD into the 20-ms MBD caused significantly lower intelligibility compared with that obtained with the preserved MBD of 20 ms ($p = 0.021$). Between the stimuli of the same MBD of 40, 80, or 160 ms, no significant differences in intelligibility were found ($p > 0.05$).

Table 2.6. Results of the preliminary experiment. Multiple comparisons of intelligibility between stimuli with the same MBDs.

Family of tests (<i>C</i>)	Asymp. Sig. (2-tailed) (* <i>p</i>)	Holm-Bonferroni (<i>C</i> - ** <i>i</i> + 1) × <i>p</i>	Holm-Bonferroni Sig. ($\alpha=0.05$)
×1 OMBD of 20 ms vs. ×0.5 OMBD of 40 ms	20-ms MBD 0.005	0.021	Significant
×4 OMBD of 20 ms vs. ×2 OMBD of 40 ms	80-ms MBD 0.075	0.225	Non-significant
×8 OMBD of 20 ms vs. ×4 OMBD of 40 ms	160-ms MBD 0.194	0.388	Non-significant
×2 OMBD of 20 ms vs. ×1 OMBD of 40 ms	40-ms MBD 0.495	0.495	Non-significant

**p* = *p*-value from multiple comparisons with the Wilcoxon Sign-rank test.

***i* = rank of the family of tests (the smallest *p*-value is the largest rank number).

2.4 Discussion

A preliminary experiment was performed in which the intelligibility of mosaiced English words was tested to find and select the range of mosaic block durations that would be suitable for English speech. As a first attempt, blocks in the speech with OMBDs of 20 and 40 ms were manipulated by compressing, preserving, or stretching them in time, but by keeping their spectral contents. Therefore, before and after manipulation, the MBDs still contained the same acoustic information. The resulting speech only differed in the speed. Compressing the blocks made the speech faster, and stretching the blocks made the speech slower. Twenty Indonesian speakers who speak English as a foreign language (EFL) participated. They typed what they had heard after three repetitions of each speech stimulus without guessing.

The first purpose of this experiment was to examine which mosaic block durations would be appropriate for English speech stimuli for the main experiment. The results showed that the highest intelligibility scores were obtained when the 20-ms or the 40-ms OMBD of the stimuli was preserved. Compressing the OMBD caused a decrease in the intelligibility of the mosaic words by about 20% for both the OMBDs. Furthermore, stretching the OMBD by a

factor of 2 did not reduce the intelligibility very much. Stretching the OMBD by a factor of 2 decreased intelligibility by only about 3% for the OMBD of 20 ms and by about 18% for the OMBD of 40 ms. However, when the OMBD was stretched by factor of 4 or more, the intelligibility of the English words decreased sharply by more than 50% for both the OMBDs compared to the preserved OMBD. These results were similar to the intelligibility scores that were obtained in the previous study with Japanese mosaic speech (Nakajima et al., 2018). That is, the intelligibility of Japanese sentences was almost perfect when the duration of the temporal segments was 40 ms or shorter. The results were also similar to those found in Kojima and Nakajima (2016) and Kojima et al. (2017), in which the intelligibility of English mosaic speech sentences was high around 80% for temporal segment duration of up to 40 ms. These temporal segment durations seem to agree with neural oscillations of 20–33 ms, which are considered to be involved in preserving phonemic intelligibility (Giraud & Poeppel, 2012; Chait et al., 2015). As a conclusion, the mosaic block durations of 20 and 40 ms turned out to be suitable to use in the main experiment.

Regarding the stimuli with the same MBDs, i.e., 20, 40, 80, or 160 ms, the English words with the same MBD of 20 ms after preservation were much more intelligible than those with an MBD of 20 ms after compression. Meanwhile, statistical comparisons between the intelligibility of mosaic speech stimuli with the same MBD of either 40, 80, or 160 ms after preserving or stretching did not result in any significant differences. The intelligibility did not change significantly even though the amount of information was different between stimuli with the OMBD of 20 ms and with the OMBD of 40 ms, in which the 20-ms OMBD conveyed more information than the 40-ms OMBD.

Although the results of this preliminary experiment showed overall lower intelligibility scores than the previous study (Kojima & Nakajima, 2016; Nakajima et al., 2018), the finding that intelligibility was highest at a block duration of 40 ms or shorter was impressive

considering that the listeners in this experiment were non-native English speakers (Indonesian speakers). In the previous study (Nakajima et al., 2018), which presented Japanese mosaic speech to native-Japanese listeners, intelligibility was almost perfect at the same mosaic block duration as found in this preliminary experiment, i.e., 40 ms or shorter. Therefore, it is important to further investigate the intelligibility of English mosaic speech by employing native-English speakers and to see whether the intelligibility would become nearly perfect as well or not. Besides that, it would be interesting as well to see the intelligibility performance from other non-native-English speakers, to see whether the same trend would appear as in this preliminary experiment or not.

The second purpose of this experiment was to examine whether all experimental procedures were in order, e.g., with regard to stimulus recording and stimulus presentation. Several limitations of the method used in this experiment were found. These may have influenced the intelligibility scores. First, the presentation levels of all stimuli were widely different, i.e., between 56 - 74 dBA (Appendix E). Some stimuli were heard loudly and some were heard softly—listening to a soft sound after listening to a loud sound can involve a shift in the listening threshold (Hood, 1950), making it difficult for the auditory system to extract the information. Therefore, the original speech sounds must be equalized before mosaicizing. Another limitation concerns the response time for the participant to answer each stimulus. Here, it was limited to just five seconds and this might have led to a deficit in the intelligibility scores. The listener may not have had enough time to think and to identify the stimulus, or may not have had enough time to type the sound. In the following main experiment, these issues were dealt with.

Chapter 3 - Main Experiment: Speech Intelligibility Comparison between Native and Non-Native Speakers of English

3.1 Purpose

The results of the preliminary experiment showed that the intelligibility of mosaic English words was highest when the mosaic block duration was 20 or 40 ms after preserving or stretching the original mosaic block duration (OMBD). The same segment durations were reported to be most suitable to convey linguistic information in the previous study with Japanese speech (Nakajima et al., 2018), but could not yet be claimed to indicate the best condition of mosaic speech. Therefore, further investigations with regard to speech-segment durations were conducted in this main experiment. Since the previous study (Nakajima et al., 2018) obtained near-perfect intelligibility when the Japanese mosaic speech was presented to native-Japanese listeners, native-English listeners were employed first in this main experiment using English mosaic speech. It was necessary to examine whether higher intelligibility would be obtained when the English mosaic speech stimuli were presented to native speakers of that language.

Indonesian listeners were employed in the preliminary investigation on English speech intelligibility. Their highest level of intelligibility was around 70% after the original speech sounds were mosaiced. This finding was very impressive considering that the listeners used English as a foreign language; even though the speech was degraded dramatically, it was still intelligible for non-native listeners when the mosaic block duration (MBD) was 40 ms or shorter after preserving the OMBDs. It was necessary to check whether the same trend on intelligibility would appear when English mosaic speech is presented to other non-native English speakers with a different language background. Therefore, in this main experiment,

two non-native English speaker groups were employed as listeners, i.e., Indonesian and Chinese speakers.

The first purpose of this main experiment was to investigate whether the intelligibility of English mosaic speech in each MBD after compressing, preserving, or stretching the OMBD of 20 and 40 ms, would show the same trend as in the preliminary experiment, but with listeners with different language backgrounds. The second purpose was to investigate whether the effects of compressing, preserving, or stretching mosaic speech would be similar among listeners with different language backgrounds. The intelligibility would be compared between the listener groups to see whether it would be the same or different between native- and non-native English listeners when the speed of English mosaic speech was preserved, speeded-up by compressing the OMBD, or slowed-down by stretching the OMBD.

3.2 Method

Compared to the preliminary experiment, two limitations were remedied in this experiment. That is, the levels of the original speech sounds were equalized first before mosaicizing in this main experiment (see Appendix F). Another change regards the response time; there was no limit of time for the participants to type what they had heard after the stimulus was presented.

3.2.1 Participants

Native speakers from three language groups, i.e., English speakers ($n = 19$; 4 speakers from Canada, 13 speakers from the United States of America, and 2 speakers from Australia; 10 males and 9 females, 20–56 years old), Indonesian speakers ($n = 19$; 9 males and 10 females, 18–42 years old), and Chinese speakers ($n = 20$; 6 males and 14 females, 22–29 years old),

participated in this experiment. The Indonesian and the Chinese participants were university students who had completed tests of English as a second language. Out of the 19 Indonesian participants, 4 had scores on TOEFL ITP (scores = 520–643), 1 had taken TOEFL iBT (score = 110), 12 had scores on the International English Language Testing System (IELTS; scores = 6.5–8.0), and 2 had taken TOEIC (scores = 720–725). From the 20 Chinese participants, 7 had scores on the Test of English as a Foreign Language (TOEFL IBT; scores = 56–89), 11 had scores on the Test of English for International Communication (TOEIC; scores = 510–880), while 2 had taken the College English Test (CET-6; scores = 543–640). For all participants, a pure-tone hearing-level test was done before the start of the experiment. All participants showed normal hearing with a loss of 30 dB or less for tones in between 250–8000 Hz, except for one English speaker (56 years old, threshold of 35–40 dB at 4000–8000 Hz, left ear). Prior to the experiment, the participants received an explanation about the procedure of the experiment. All agreed to participate and provided written informed consent. The experiment was conducted with prior approval of the Ethics Committee of Kyushu University. The present thesis research was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Kyushu University (Project Identification code: 457, the approval code was 70-6).

3.2.2 Equipment

Stimulus recording

In this experiment, the same equipment as in the preliminary experiment was used (see Section 2.2.2).

Stimulus presentation

With regard to stimulus presentation, for the native-English and the Indonesian participants, the equipment was set as follows. The experiment was conducted in different room conditions, and the background noise varied mostly in between 30–45 dBA. The stimuli were stored in a computer notebook (Toshiba, Dynabook R734, Tokyo, Japan). From the computer, the stimuli were passed through a USB headphone amplifier (Audio-Technica, AT-HA40USB, Tokyo, Japan), before being presented to the listener via headphones (Roland, RH-300, Hamamatsu, Japan). All stimuli were presented at a presentation level of 66–75 dBA (Fast-Peak), as measured with a sound level meter (ACO, Type 6240).

Meanwhile, for the Chinese participants, the experiment was conducted in the same soundproof room as where the stimulus recording had taken place with the same equipment as in the preliminary experiment.

3.2.3 Stimuli

English word specifications

Since the intelligibility scores of English words obtained from native-English and non-native English listeners were compared, the words used as stimuli in this experiment were replaced by words with a higher difficulty level than the difficulty level of the words used in the preliminary experiment. Whereas textbooks for children (up to elementary school) were used in the preliminary experiment, a textbook for teachers was used in this experiment. Eighty English words in a Consonant-Vowel-Consonant (CVC) structure were used. A reason for being interested in CVCs was that this thesis research hoped to pick up some information on which consonants (specifically as the initial consonant) are easy (or difficult) to be identified when mosaiced. The words were derived from an English textbook (Kress & Fry, 2016) within the category of “content words”, which have lexical meanings (Richards & Schmidt,

2010). The words were selected as follows. First, we applied some criteria to avoid any ambiguity in word meaning. Content words ending with the letter “r” were not considered, since its pronunciation can sometimes be lost, which is known as a non-rhotic accent (e.g., four; [fɔ:(r)]; (Carley et al., 2018). Furthermore, words with two or more possible pronunciations depending on dialect were excluded (e.g., dog; [dɒg/] or [dɔg/], (Wells, 2008). Finally, words that appeared in a homophone or heteronym list were excluded as well (Richards & Schmidt, 2010). Based on the criteria, 109 words were collected. Following this, each word was presented to five native-English speakers in order to check whether it could be easily understood shortly after hearing. If not, it was omitted, else, we further checked the phonetic pronunciation of the words and clustered them according to 18 initial consonants (/p/, /b/, /t/, /d/, /k/, /g/, /s/, /ʃ/, /tʃ/, /dʒ/, /f/, /h/, /m/, /n/, /l/, /r/, /w/, /j/), 10 vowels (/æ/, /ɪ/, /ʊ/, /e/, /ɪ/, /i:/, /u:/, /aɪ/, /aʊ/, /eɪ/), and 18 final consonants (/p/, /b/, /t/, /d/, /k/, /g/, /s/, /z/, /ʃ/, /tʃ/, /dʒ/, /f/, /v/, /θ/, /m/, /n/, /ŋ/, /l/).

To further reduce the 109 words into the final 80 words, the following steps were taken. First, we selected words so that all phonetic categories were represented in the list. Second, we checked their intelligibility according to the results of the preliminary experiment for Indonesian listeners (Chapter 1; Santi et al., 2019), and omitted words that were not intelligible. For example, we selected words with a vowel /aɪ/ and last consonant /v/ (e.g., five and dive), because they were fairly intelligible at about 80% (Santi et al., 2019). Words with a vowel /ʊ/ and last consonant /l/ (e.g., “full” and “pull”) were omitted, because they were unintelligible, with an intelligibility of about 10% (Santi et al., 2019). Finally, we checked whether the words were included in the top 1000, 2000, or 3000 most frequently used words in both spoken and written English (Wells., 2014). For example, the words “fish” and “rush” were included in the final stimulus list, but the words “shed” and “hedge” were not.

As a result, there were four or five words for each initial consonant on the list, except for the consonant /j/, for which only three words were used. The selected words then were divided into 20 groups. In order to generate sound variety, each initial consonant, vowel, and last consonant appeared once only in each group (see Table 3.1). Ten other words were used for practice trials, taken from the omitted words, and chosen randomly. These practice trials introduced the stimulus types.

Table 3.1. The 80 CVC-words used in the main experiment about English mosaic speech.

Group	Word	Group	Word	Group	Word	Group	Word
A	bush /bʊʃ/	F	tab /tæb/	K	push /pʊʃ/	P	cave /keɪv/
	love /lʌv/		rule /ru:l/		dive /daɪv/		tell /tel/
	rage /reɪdʒ/		line /laɪn/		juice /dʒu:s/		peach /pi:tʃ/
	feed /fi:d/		sheep /ʃi:p/		gun /gʌn/		hat /hæt/
B	rush /rʌʃ/	G	wife /waɪf/	L	king /kɪŋ/	Q	wish /wɪʃ/
	date /deɪt/		soup /su:p/		touch /tʌtʃ/		cheese /tʃi:z/
	nine /naɪn/		big /bɪg/		nap /næp/		game /geɪm/
	food /fu:d/		couch /kaʊtʃ/		move /mu:v/		young /jʌŋ/
C	size /saɪz/	H	pig /pɪg/	M	name /neɪm/	R	book /bʊk/
	yell /jel/		cook /kʊk/		lab /læb/		keep /ki:p/
	fish /fɪʃ/		shape /ʃeɪp/		guide /gaɪd/		safe /seɪf/
	mouse /maʊs/		gel /dʒel/		chief /tʃi:f/		nice /naɪs/
D	tag /tæg/	I	rub /rʌb/	N	youth /ju:θ/	S	hang /hæŋ/
	beep /bi:p/		mess /mes/		hate /heɪt/		wise /waɪz/
	shut /ʃʌt/		give /gɪv/		deep /di:p/		loud /laʊd/
	wing /wɪŋ/		wake /weɪk/		rise /raɪz/		june /dʒu:n/
E	tooth /tu:θ/	J	dish /dɪʃ/	O	head /hed/	T	south /saʊθ/
	doubt /daʊt/		mood /mu:d/		gum /gʌm/		face /feɪs/
	check /tʃek/		judge /dʒʌdʒ/		shine /ʃaɪn/		map /mæp/
	page /peɪdʒ/		five /faɪv/		choose /tʃu:z/		life /laɪf/

Stimulus recording

All words were pronounced by a male, native-English speaker (from the United States of America, age = 28 years old). The original speech recordings were stored in the same file format as in the preliminary experiment, following the same procedures (Appendices A and B). All original speech sounds were equalized in intensity before being mosaiced and used as stimuli (Appendix F).

English speech mosaicization

The English speech sounds were mosaicized with the same procedure as described in Section 2.2.3.

3.2.4 Procedures

The experiment was divided into two sessions for each language group. The first session was for mosaic speech stimuli. The 80 CVC words were divided into twenty groups, each containing four words (Table 3.1). Each group was assigned to a different mosaic speech stimulus type, and this assignment was different among participants (see Table 2.3). All participants received all the words but in different stimulus types. The 10 words that were used for practice trials were also randomly assigned to a different stimulus type, and the assignment was the same among all participants. There were 5–10 minutes as a break before the next session. In the second session, all original speech stimuli were presented to all participants. Both sessions started with one block of practice trials and were followed by four main blocks, each containing twenty measurement trials.

In each session, the stimuli were presented through headphones in random order to the participant, who sat on a chair in front of the computer interface, which was created on Visual Basic .NET programming language (Visual Studio 2019 version 16.0). The participant was asked to click a “play” button on the interface to start a trial. The stimulus of each trial was presented 0.5 s after the button was clicked. The presentation was repeated three times with 1.5-s intervals. After listening to the sound stimulus, the participant typed the perceived word, if any, using the English alphabet. The participant was instructed to avoid guessing the correct answer. There was no limited time for the participant to respond to each stimulus, but the time needed to respond was recorded.

3.2.5 Statistical Analysis

The same statistical methods were applied to the intelligibility scores as in the preliminary experiment for each participant group, and the results were compared to each other (Appendix H).

3.3 Results

3.3.1 Intelligibility Comparisons between Original Speech and Mosaic Speech

Figure 3.1 shows how many words the participants identified correctly when presented in their original form. The native-English group performed almost perfectly, while the Indonesian participants scored close to 90% correct and the Chinese participants scored about 80% correct.

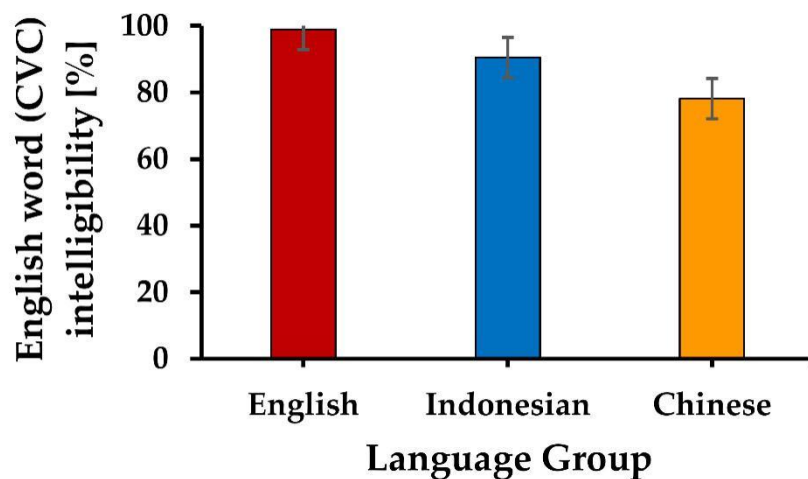


Figure 3.1. Results of the main experiment. Word identification accuracy (intelligibility) for original speech for each language group (English, $n = 19$; Indonesian, $n = 19$; Chinese, $n = 20$). Error bars indicate standard error of means.

The intelligibility scores for the mosaic speech stimuli for the three language groups are shown in Figure 3.1. Since the results of the Wilcoxon signed-rank test (Field, 2009) had

shown that there was no significant effect of half-phase or whole-phase starting phases on word intelligibility ($p > 0.05$), and since investigating the effects of phase was not the main purpose of this present thesis research, for convenience the scores were collapsed for subsequent analyses. Figure 3.1 shows that mosaicing the original speech really affected the intelligibility. The intelligibility decreased by about 21% for the native-English group, by about 40% for the Indonesian group, and by about 25% for the Chinese group in comparison to the original speech stimuli. An obvious reason for this is that when original speech is mosaiced, its signal degrades both in the temporal and the frequency dimension.

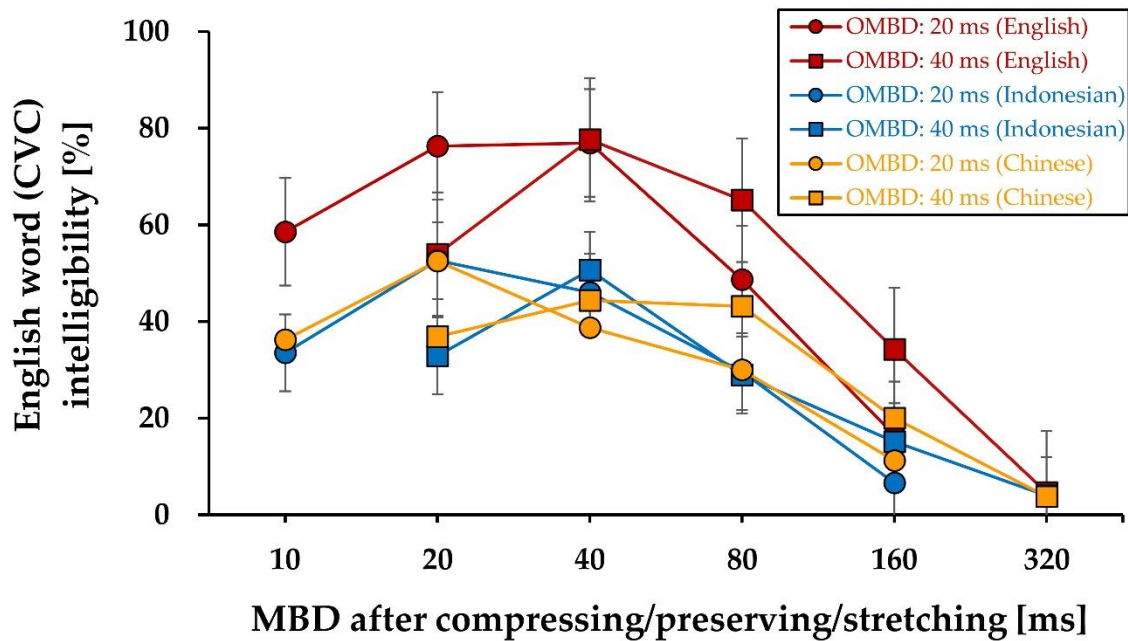


Figure 3.2. Results of the main experiment. English word identification accuracy (intelligibility) for mosaic speech as functions of MBD after compressing/preserving/stretching (English, $n = 19$; Indonesian, $n = 19$; Chinese, $n = 20$). The data for the half-phase and the whole-phase types are collapsed. Error bars indicate standard errors.

3.3.2 The Effect of Compressing or Stretching the Original Mosaic Block Duration (OMBD)

Native-English listeners

The intelligibility of English mosaic speech for the native-English listeners was highest at the MBD of 20 ms after preserving and at the MBD of 40 ms after preserving or stretching, and decreased as the MBD was shorter or longer (see Figure 3.2). When the OMBDs were compressed, the intelligibility decreased by about 20% from the intelligibility of the preserved OMBD. When the OMBDs were stretched into an MBD of 80 ms, the intelligibility decreased sharply. For an MBD of 320 ms after stretching, the intelligibility was close to zero.

A Friedman two-way analysis of variance by ranks (Field, 2009) was performed to analyze the main effects of compressing or stretching the OMBD. There was a significant difference in intelligibility between mosaic speech types ($n = 19$, $k = 19$, $\chi^2 = 188.004$; $p < 0.001$). Multiple comparisons with the Wilcoxon Signed-rank test (Field, 2009) were used to follow up this finding, and a Holm-Bonferroni correction (Holm, 1979; Abdi, 2010) was performed after that. In detail (see Table 3.2), compressing the OMBD deteriorated intelligibility (OMBD of 20 ms: $p = 0.010$; OMBD of 40 ms: $p = 0.009$). The intelligibility of the MBD after preserving was significantly higher than that of the MBD after stretching (80, 160, or 320 ms; $p < 0.05$). Stretching the OMBD by a factor of 2 led to significantly higher intelligibility than compressing the OMBD by half for the OMBD of 20 ms ($p = 0.010$), but not for the OMBD of 40 ms ($p = 0.075$). Moreover, stretching the OMBD by a factor of 2 induced higher intelligibility than stretching the OMBD by a factor of 4 or 8 ($p < 0.01$).

Table 3.2. Results of the main experiment. Multiple comparisons of intelligibility between compressed, preserved, and stretched mosaic speech conditions within the OMBD of 20 or 40 ms for the native-English listeners (n=19).

	Family of tests (<i>C</i>)	Rank (<i>i</i>)	Asymp. Sig. (2-tailed) (* <i>p</i>)	Holm-Bonferroni (<i>C - i + 1</i>) × <i>p</i>	Holm-Bonferroni Sig. ($\alpha = 0.05$)
OMBD of 20 ms:					
×1 vs. ×8	20 and 160-ms MBD	1	0.000	0.001	Significant (×1 > ×8)
×2 vs. ×8	40 and 160-ms MBD	2	0.000	0.001	Significant (×2 > ×8)
×0.5 vs. ×8	10 and 160-ms MBD	3	0.000	0.001	Significant (×0.5 > ×8)
×1 vs. ×4	20 and 80-ms MBD	4	0.000	0.003	Significant (×1 > ×4)
×2 vs. ×4	40 and 80-ms MBD	5	0.001	0.003	Significant (×2 > ×4)
×4 vs. ×8	80 and 160-ms MBD	6	0.001	0.005	Significant (×4 > ×8)
×0.5 vs. ×2	10 and 40-ms MBD	7	0.003	0.010	Significant (×0.5 < ×2)
×0.5 vs. ×1	10 and 20-ms MBD	8	0.003	0.010	Significant (×0.5 < ×1)
×0.5 vs. ×4	10 and 80-ms MBD	9	0.096	0.192	Non-significant
×1 vs. ×2	20 and 40-ms MBD	10	0.812	0.812	Non-significant
OMBD of 40 ms:					
×1 vs. ×4	40 and 160-ms MBD	1	0.000	0.001	Significant (×1 > ×4)
×2 vs. ×4	80 and 160-ms MBD	2	0.000	0.002	Significant (×2 > ×4)
×0.5 vs. ×1	20 and 40-ms MBD	3	0.002	0.009	Significant (×0.5 < ×1)
×0.5 vs. ×4	20 and 160-ms MBD	4	0.005	0.016	Significant (×0.5 > ×4)
×1 vs. ×2	40 and 80-ms MBD	5	0.008	0.017	Significant (×1 > ×2)
×0.5 vs. ×2	20 and 80-ms MBD	6	0.075	0.075	Non-significant

**p* = *p*-value from multiple comparisons with the Wilcoxon Sign-rank test

Non-native English listeners

Indonesian listeners

The intelligibility of English mosaic speech for the Indonesian listeners was highest for the MBDs of 20 or 40 ms after preserving and decreased as the MBD was shorter or longer, similar to the intelligibility data of the native-English listeners. The intelligibility decreased by about 19% for the MBDs after compressing the OMBDs. The intelligibility decreased sharply when the OMBDs were stretched into MBDs of 80 ms or longer.

By the same statistical method as described previously, here too, significant differences in intelligibility between mosaic speech types were found ($n = 19, k = 19, \chi^2 = 153.618, p < 0.001$). In detail (see Table 3.3), the intelligibility of mosaic speech with the MBD of 20 ms

after preserving was significantly higher than that of the MBD after compressing ($p = 0.007$), but the same significant difference did not appear for the MBDs of 40 ms after preserving and after compressing ($p = 0.077$). The intelligibility of the MBDs after preserving was significantly higher than the MBDs after stretching (80, 160, or 320 ms; $p < 0.01$). The higher intelligibility was obtained for the stimuli with the MBD of 40 ms ($p < 0.005$) or 80 ms ($p < 0.05$) after stretching compared with the stimuli with the MBD of 160 ms after stretching.

Although the Indonesian listeners overall had significantly lower intelligibility scores than the native-English listeners, as can be seen in Figure 3.2, there were similar trends in intelligibility across the two language groups.

Table 3.3. Results of the main experiment. Multiple comparisons of intelligibility between compressed, preserved, and stretched mosaic speech conditions within the OMBD of 20 or 40 ms for the Indonesian listeners (n=19).

Family of tests (<i>C</i>)	Rank (<i>i</i>)	Asymp. Sig. (2-tailed) (* <i>p</i>)	Holm-Bonferroni (<i>C - i + 1</i>) × <i>p</i>	Holm-Bonferroni Sig. ($\alpha = 0.05$)	
OMBD of 20 ms:					
×1 vs. ×8	20 and 160-ms MBD	1	0.000	0.001	Significant (×1 > ×8)
×0.5 vs. ×8	10 and 160-ms MBD	2	0.000	0.002	Significant (×0.5 > ×8)
×2 vs. ×8	40 and 160-ms MBD	3	0.000	0.003	Significant (×2 > ×8)
×1 vs. ×4	20 and 80-ms MBD	4	0.001	0.004	Significant (×1 > ×4)
×0.5 vs. ×1	10 and 20-ms MBD	5	0.001	0.007	Significant (×0.5 < ×1)
×4 vs. ×8	80 and 160-ms MBD	6	0.003	0.014	Significant (×4 > ×8)
×2 vs. ×4	40 and 80-ms MBD	7	0.052	0.209	Non-significant
×0.5 vs. ×2	10 and 40-ms MBD	8	0.109	0.326	Non-significant
×1 vs. ×2	20 and 40-ms MBD	9	0.346	0.692	Non-significant
×0.5 vs. ×4	10 and 80-ms MBD	10	0.382	0.382	Non-significant
OMBD of 40 ms:					
×1 vs. ×4	40 and 160-ms MBD	1	0.000	0.001	Significant (×1 > ×4)
×1 vs. ×2	40 and 80-ms MBD	2	0.002	0.008	Significant (×1 > ×2)
×0.5 vs. ×4	20 and 160-ms MBD	3	0.009	0.036	Significant (×0.5 > ×4)
×2 vs. ×4	80 and 160-ms MBD	4	0.015	0.046	Significant (×2 > ×4)
×0.5 vs. ×1	20 and 40-ms MBD	5	0.038	0.077	Non-significant
×0.5 vs. ×2	20 and 80-ms MBD	6	0.732	0.732	Non-significant

**p* = *p*-value from multiple comparisons with the Wilcoxon Sign-rank test

Chinese listeners

The intelligibility of English mosaic speech for the Chinese listeners was highest as well for the MBDs of 20 ms or 40 ms after preserving and decreased as the MBD was shorter or longer, similar to the intelligibility data of the native-English and the Indonesian listeners. The intelligibility decreased by about 8-16% for the MBDs after compressing compared to the intelligibility of the preserved OMBD. The intelligibility decreased sharply when the OMBDs were stretched into MBDs of 80 ms or longer, except when the OMBD of 40 ms was stretched by a factor of 2.

For the Chinese participants, significant differences also were found in intelligibility between mosaic speech types ($n = 20$, $k = 19$, $\chi^2 = 129.139$, $p < 0.001$). In detail (see Table 3.4), for both OMBDs, there were no significant differences in intelligibility between the preserved and the compressed or the stretched $\times 2$ stimuli ($p > 0.05$), nor between the compressed and the stretched $\times 2$ or $\times 4$ stimuli ($p > 0.05$). The Chinese group obtained higher intelligibility for the stimuli with the MBD of 40 ms ($p < 0.005$) or 80 ms ($p < 0.05$) after stretching compared with the stimuli with the MBD of 160 ms after stretching.

Table 3.4. Results of the main experiment. Multiple comparisons of intelligibility between compressed, preserved, and stretched mosaic speech conditions within the OMBD of 20 or 40 ms for the Chinese listeners (n=20).

Family of tests (<i>C</i>)	Rank (<i>i</i>)	Asymp. Sig. (2-tailed) (* <i>p</i>)	Holm-Bonferroni (<i>C</i> - <i>i</i> + 1) × <i>p</i>	Holm-Bonferroni Sig. ($\alpha=0.05$)
OMBD of 20 ms:				
×1 vs. ×8 20 and 160-ms MBD	1	0.000	0.002	Significant (×1 > ×8)
×2 vs. ×8 40 and 160-ms MBD	2	0.000	0.003	Significant (×2 > ×8)
×0.5 vs. ×8 10 and 160-ms MBD	3	0.000	0.003	Significant (×0.5 > ×8)
×4 vs. ×8 80 and 160-ms MBD	4	0.002	0.016	Significant (×4 > ×8)
×1 vs. ×4 20 and 80-ms MBD	5	0.003	0.018	Significant (×1 > ×4)
×0.5 vs. ×1 10 and 20-ms MBD	6	0.013	0.064	Non-significant
×1 vs. ×2 20 and 40-ms MBD	7	0.017	0.067	Non-significant
×2 vs. ×4 40 and 80-ms MBD	8	0.113	0.340	Non-significant
×0.5 vs. ×4 10 and 80-ms MBD	9	0.328	0.656	Non-significant
×0.5 vs. ×2 10 and 40-ms MBD	10	0.711	0.711	Non-significant
OMBD of 40 ms:				
×2 vs. ×4 80 and 160-ms MBD	1	0.001	0.007	Significant (×2 > ×4)
×1 vs. ×4 40 and 160-ms MBD	2	0.003	0.013	Significant (×1 > ×4)
×0.5 vs. ×4 20 and 160-ms MBD	3	0.014	0.056	Non-significant
×0.5 vs. ×1 20 and 40-ms MBD	4	0.186	0.557	Non-significant
×0.5 vs. ×2 20 and 80-ms MBD	5	0.292	0.583	Non-significant
×1 vs. ×2 40 and 80-ms MBD	6	0.982	0.982	Non-significant

**p*= *p*-value from multiple comparisons with the Wilcoxon Sign-rank test

Although the Chinese listeners overall had significantly lower intelligibility scores than the native-English and the Indonesian listeners, as can be seen in Figure 3.2, there were similar trends in intelligibility across the three language groups.

3.3.3 Intelligibility Comparisons between Stimuli with the Same MBDs within Each Language Group

Due to stretching or compressing, the mosaic speech stimuli with a 20- and a 40-ms OMBD contained blocks of the same duration, and it was important to analyze whether the word intelligibility for stimuli with the same MBD would be similar or not. The results are shown in Table 3.5.

Native-English listeners

Regarding the stimuli with an MBD of 20 ms, compressing the 40-ms OMBD into the 20-ms MBD caused significantly lower intelligibility compared with that obtained with the MBD of 20 ms ($p = 0.004$) after preserving. Between the stimuli of the same MBD of 40, 80, or 160 ms, no significant differences in intelligibility were found ($p > 0.05$).

Non native-English listeners

For the Indonesian group, there was no significant effect of compression for the stimuli with an OMBD of 20 ms ($p = 0.088$). Moreover, the Indonesian group had similar results as the native-English group, in that there were no significant differences in intelligibility between the stimuli with an MBD of 40 ms, 80 ms, or 160 ms ($p > 0.05$). Meanwhile, for the Chinese group, similar to the native- English group, compressing the 40-ms OMBD into 20-ms blocks caused significantly lower intelligibility compared with that obtained with the MBD of 20 ms after preserving ($p = 0.002$). Furthermore, for the Chinese group, there were significant differences between the stimuli with an OMBD of 20 ms stretched into 80 ms and the stimuli with the OMBD of 40 ms stretched into 80 ms ($p < 0.05$). Similarly, there were also significant differences between the stimuli with an OMBD of 20 ms stretched into 160 ms and the stimuli with an OMBD of 40 ms stretched into 160 ms ($p < 0.05$).

Table 3.5. Results of the main experiment. Multiple comparisons of intelligibility between stimuli with the same MBDs within each language group.

Family of tests (C)		Asymp. Sig. (2-tailed) (*p)	Holm-Bonferroni (C - **i + 1) × p	Holm-Bonferroni Sig. (α= 0.05)
Native-English				
×1 OMBD of 20 ms vs. ×0.5 OMBD of 40 ms	20-ms MBD	0.004	0.015	Significant (×1 OMBD of 20 ms > ×0.5 OMBD of 40 ms)
×4 OMBD of 20 ms vs. ×2 OMBD of 40 ms	80-ms MBD	0.023	0.069	Non-significant
×8 OMBD of 20 ms vs. ×4 OMBD of 40 ms	160-ms MBD	0.111	0.222	Non-significant
×2 OMBD of 20 ms vs. ×1 OMBD of 40 ms	40-ms MBD	0.417	0.417	Non-significant
Indonesian				
×1 OMBD of 20 ms vs. ×0.5 OMBD of 40 ms	20-ms MBD	0.022	0.088	Non-significant
×8 OMBD of 20 ms vs. ×4 OMBD of 40 ms	160-ms MBD	0.096	0.289	Non-significant
×2 OMBD of 20 ms vs. ×1 OMBD of 40 ms	40-ms MBD	0.548	1.096	Non-significant
×4 OMBD of 20 ms vs. ×2 OMBD of 40 ms	80-ms MBD	0.979	0.979	Non-significant
Chinese				
×1 OMBD of 20 ms vs. ×0.5 OMBD of 40 ms	20-ms MBD	0.002	0.008	Significant (×1 OMBD of 20 ms > ×0.5 OMBD of 40 ms)
×8 OMBD of 20 ms vs. ×4 OMBD of 40 ms	160-ms MBD	0.005	0.016	Significant (×8 OMBD of 20 ms < ×4 OMBD of 40 ms)
×4 OMBD of 20 ms vs. ×2 OMBD of 40 ms	80-ms MBD	0.023	0.047	Significant (×4 OMBD of 20 ms < ×2 OMBD of 40 ms)
×2 OMBD of 20 ms vs. ×1 OMBD of 40 ms	40-ms MBD	1	1	Non-significant

*p= p-value from multiple comparisons with the Wilcoxon Sign-rank test

**i= rank tests

3.3.4 Intelligibility Comparisons between the Language Groups

One of the present thesis research purposes was to investigate whether the intelligibility of mosaic speech would differ between the native-English and the non-native English listeners (Indonesian and Chinese listeners).

Table 3.6. Results of the main experiment. Intelligibility comparisons between the language groups.

Family of tests (<i>C</i>)		Asymp. Sig. (2-tailed) (* <i>p</i>)	Holm-Bonferroni (<i>C</i> - ** <i>i</i> + 1) × <i>p</i>	Holm-Bonferroni Sig. ($\alpha=0.05$)
OMBD of 20 ms:				
10-ms MBD	English vs. Indonesian	0.002	0.005	Significant (English > Indonesian)
	English vs. Chinese	0.002	0.003	Significant (English > Chinese)
	Indonesian vs. Chinese	0.774	0.774	Non-significant
20-ms MBD	English vs. Indonesian	0.001	0.002	Significant (English > Indonesian)
	English vs. Chinese	0.002	0.004	Significant (English > Chinese)
	Indonesian vs. Chinese	0.796	0.796	Non-significant
40-ms MBD	English vs. Chinese	0.000	0.001	Significant (English > Chinese)
	English vs. Indonesian	0.001	0.001	Significant (English > Indonesian)
	Indonesian vs. Chinese	0.296	0.296	Non-significant
80-ms MBD	English vs. Chinese	0.016	0.047	Significant (English > Chinese)
	English vs. Indonesian	0.019	0.038	Significant (English > Indonesian)
	Indonesian vs. Chinese	0.931	0.931	Non-significant
160-ms MBD	Indonesian vs. English	0.041	0.124	Non-significant
	English vs. Chinese	0.210	0.421	Non-significant
	Indonesian vs. Chinese	0.264	0.264	Non-significant
OMBD of 40 ms:				
20-ms MBD	English vs. Chinese	0.002	0.005	Significant (English > Chinese)
	English vs. Indonesian	0.002	0.004	Significant (English > Indonesian)
	Indonesian vs. Chinese	0.832	0.832	Non-significant
40-ms MBD	English vs. Chinese	0.000	0.001	Significant (English > Chinese)
	English vs. Indonesian	0.001	0.001	Significant (English > Indonesian)
	Indonesian vs. Chinese	0.253	0.253	Non-significant
80-ms MBD	English vs. Indonesian	0.000	0.001	Significant (English > Indonesian)
	English vs. Chinese	0.001	0.001	Significant (English > Chinese)
	Indonesian vs. Chinese	0.011	0.011	Significant (Indonesian < Chinese)
160-ms MBD	English vs. Indonesian	0.001	0.004	Significant (English > Indonesian)
	English vs. Chinese	0.011	0.022	Significant (English > Chinese)
	Indonesian vs. Chinese	0.144	0.144	Non-significant
320-ms MBD	English vs. Indonesian	0.931	2.794	Non-significant
	English vs. Chinese	0.931	1.863	Non-significant
	Indonesian vs. Chinese	1.000	1	Non-significant

**p*= *p*-value from multiple comparisons with the Wilcoxon Sign-rank test

***i*= rank tests

The intelligibility (word accuracy) of the native-English group was higher than that of the Indonesian group or the Chinese group in any comparisons ($p < 0.05$), except at an MBD of 160 ms for stimuli with a 20-ms OMBD ($p= 0.27$ between the Indonesian group; $p= 0.44$ between the Chinese group). Meanwhile, the intelligibility comparisons between the Indonesian and the Chinese group did not differ significantly ($p > 0.05$) in any of the same

MBD durations in both OMBDs. In sum, the intelligibility was almost similar between the Indonesian and the Chinese group, while, as expected, the English participants showed the highest intelligibility in any preserved/stretched MBD for both OMBDs.

3.3.5 Intelligibility of Phonemes of the Initial Consonant of Each English Word

As mentioned in Section 2.2.3, there were two types of mosaicing phase, i.e., the half-phase type and the whole-phase type. The purpose of this was to explore whether phoneme perception would be affected if the mosaicing began a half-block duration or one block duration earlier than the onset of the speech. There were 18 initial consonants (/p/, /b/, /t/, /d/, /k/, /g/, /s/, /ʃ/, /tʃ/, /dʒ/, /f/, /h/, /m/, /n/, /l/, /r/, /w/, /j/; Section 3.2.3), which each appeared four or five times on the list of English words, except for the consonant /j/, which only three times appeared.

The intelligibility of the phonemes was obtained by counting the number of correct phonemes given by the participants for the initial consonant in each word, both when the whole word was answered correctly and when not. Five consonant categories were made: stop consonants (/b/, /d/, /g/, /p/, /t/, /k/), fricatives (/f/, /s/, /ʃ/, /h/), approximants (/w/, /j/, /l/, /r/), nasals (/m/, /n/), and affricates (/tʃ/, /dʒ/) (Ladefoged, 2005). As a note, the phoneme was considered intelligible when its intelligibility score was 50% or above. Since there was no effect of mosaicing phase types on phoneme perception of English words (as mentioned in Section 3.3.1), the scores were collapsed in the following summaries for each group (the native-English, the Indonesian, and the Chinese groups).

Native-English listeners

For the stop consonants, the phoneme /b/ was intelligible only when the OMBDs were compressed by half. Furthermore, the phonemes /d/ and /t/ were intelligible from the MBDs of 10-40 ms, the phoneme /p/ was intelligible at the MBDs of 20 and 40 ms, and the phoneme /k/ was intelligible from the MBDs of 10-80 ms, after the OMBDs were compressed or preserved or stretched. Moreover, the phoneme /g/ was intelligible at the MBDs of 20-80 ms after the OMBDs were preserved or stretched.

For the fricative consonants, all phonemes were intelligible from the MBDs of 10-80 ms after the OMBD of 20 ms was compressed or preserved or stretched, except for the phoneme /h/, which was unintelligible at an MBD of 80 ms. The phoneme /s/ was intelligible overall in any MBDs of the OMBD of 20 ms. Meanwhile, at the OMBD of 40 ms, the phonemes /f/ and /ʃ/ were intelligible from the MBDs of 20-160 ms after the OMBD was compressed or preserved, or stretched. For the OMBD of 40 ms, the phonemes /ʃ/ and /h/ were intelligible at the MBDs of 40 and 80 ms after the OMBD was preserved or stretched.

For the approximant consonants, all phonemes in this category were intelligible from the MBDs of 10-160 ms after the OMBDs were compressed or preserved or stretched, except for the phonemes /w/ and /l/ at the MBD of 160 ms after stretching the OMBD of 20 ms. Furthermore, all phonemes in the nasal consonant category were intelligible from the MBDs of 10-160 ms. Finally, all the phonemes of affricate consonants were intelligible from the MBDs of 10-80 ms (except for the phoneme /tʃ/ when the OMBDs were compressed), after the OMBDs were compressed or preserved or stretched.

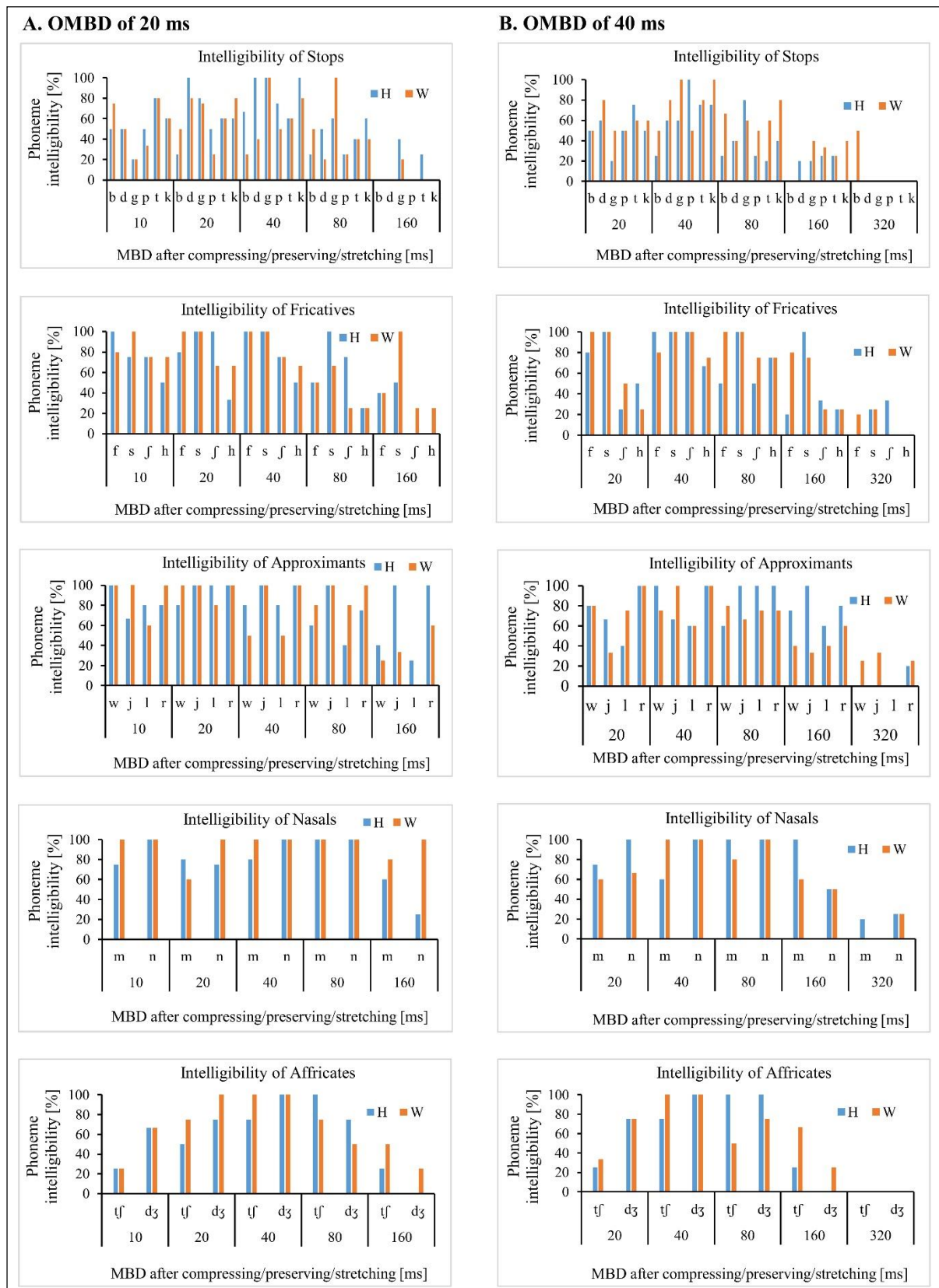


Figure 3.3. Results of the main experiment. The intelligibility of phonemes of English words as functions of MBD after compressing/preserving/stretching the OMBD of 20 ms (A) and the OMBD of 40 ms (B) for the native-English group (n=19). H indicates the half-phase type and W indicates the whole-phase type of mosaicing phase.

Non native-English listeners

In general, both the Indonesian and the Chinese groups showed lower intelligibility of the phonemes compared with that obtained in the native-English group. For the Indonesian group, for the stop consonants, the phoneme /d/ was more intelligible compared with other phonemes from the MBDs of 10-40 ms. For the fricative consonants, the phoneme /f/ was intelligible from the MBDs of 10-80 ms for the OMBD of 20 ms and from the MBDs of 40-80 ms for the OMBD of 40 ms. The phoneme /ʃ/ was intelligible when the OMBDs were preserved or stretched by a factor of 2, and the phoneme /s/ was intelligible when the OMBDs were compressed or preserved or stretched by a factor of 2 or 4. Meanwhile, for the phonemes at the OMBD of 40 ms, the phonemes /f/ and /ʃ/ were intelligible at the MBDs of 40 and 80 ms, and the phoneme /s/ was intelligible from the MBDs of 20-160 ms. For the approximant consonants, the phonemes /w/, /j/, and /r/ were intelligible from the MBDs of 10-80 ms. Furthermore, all phonemes in the nasal consonant category were also intelligible from the MBDs of 20-80 ms. Finally, all the affricate consonants were intelligible when the OMBDs were preserved (for details, see Figure 3.4).

For the Chinese group, most phonemes in the stop consonants category were intelligible when the OMBDs were preserved only, except for the phoneme /d/, which was intelligible from the MBDs of 10-40 ms. For the fricative consonants, the phonemes /f/ and /ʃ/ were intelligible from the MBDs of 10-80 ms, and the phoneme /s/ was intelligible from 10-160 ms. The phonemes /w/, /j/, and /r/ in the approximant consonants category were intelligible when the OMBDs were compressed or preserved or stretched by a factor of 2 or 4. Furthermore, all phonemes in the nasal consonant category were also intelligible from the preserved OMBDs up to 160-ms MBD. Finally, all the affricate consonants were intelligible when the OMBDs were preserved or stretched by a factor of 2 (for details, see Figure 3.5).

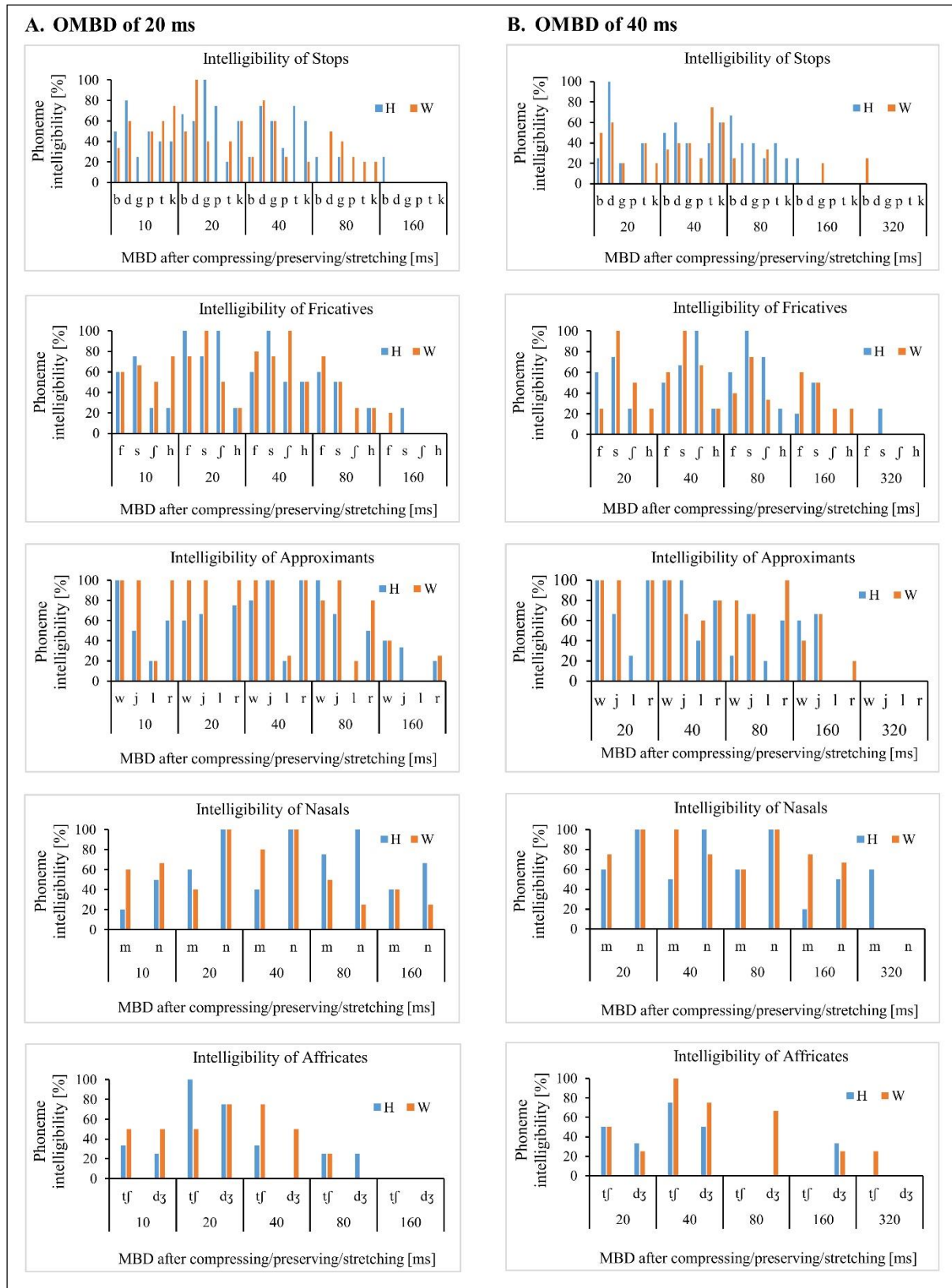


Figure 3.4. Results of the main experiment. The intelligibility of phonemes of English words as functions of MBD after compressing/preserving/stretching the OMBD of 20 ms (A) and the OMBD of 40 ms (B) for the Indonesian group ($n=19$). H indicates the half-phase type and W indicates the whole-phase type of mosaicing phase.

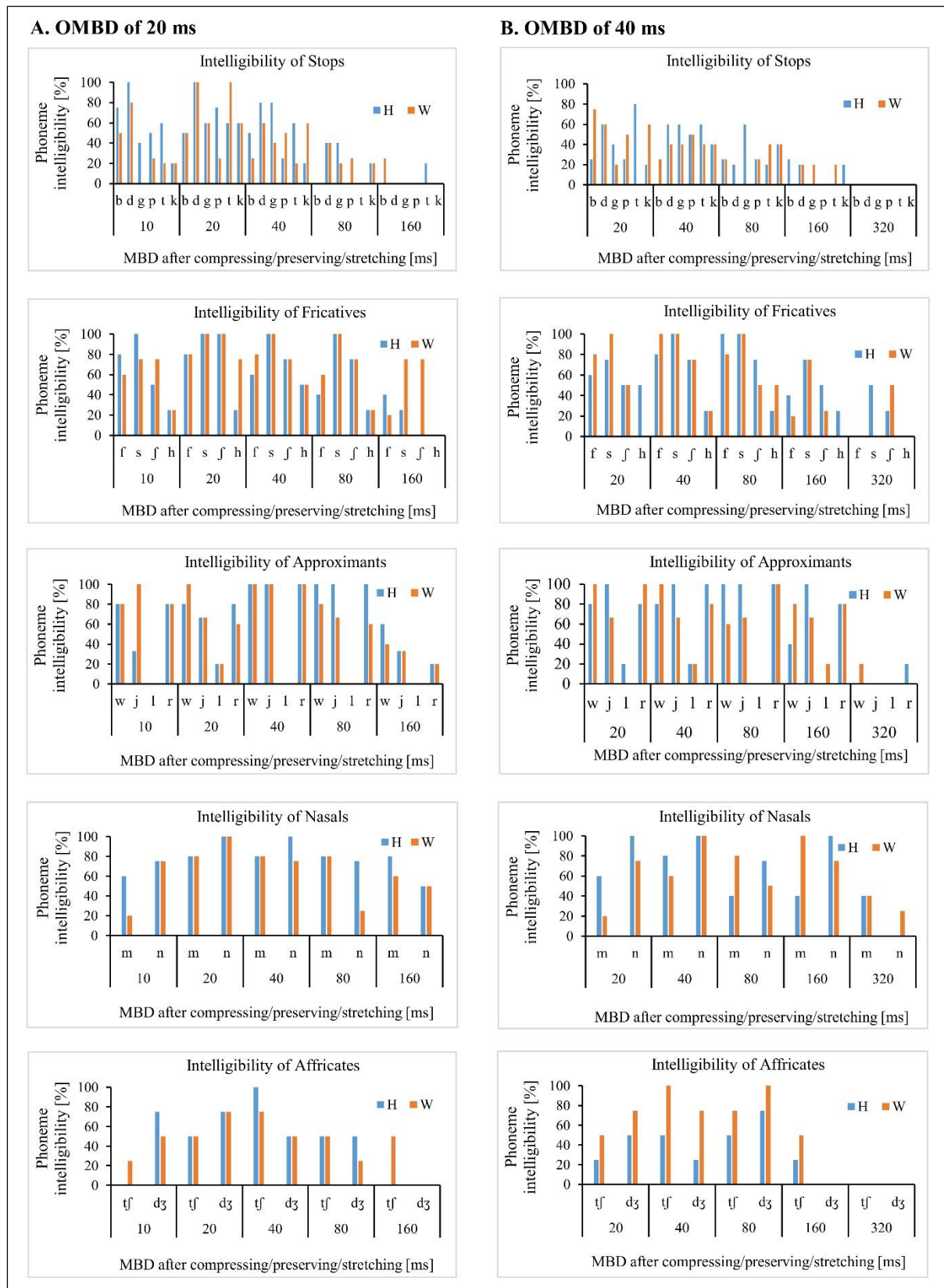


Figure 3.5. Results of the main experiment. The intelligibility of phonemes of English words as functions of MBD after compressing/preserving/stretching the OMBD of 20 ms (A) and the OMBD of 40 ms (B) for the Chinese group (n=20). H indicates the half-phase type and W indicates the whole-phase type of mosaicing phase.

3.4 Discussion

Listeners with three different language backgrounds (native-English, Indonesian, and Chinese) were employed in this main experiment to determine the intelligibility of mosaiced English words; they typed what they had heard. The results showed that, overall, for each language group, the highest intelligibility was obtained when the OMBD of the stimuli was preserved/stretched into 20-ms or 40-ms MBDs (intelligibility scores for the English group: 76–78%; for the Indonesian group: 50–53%; and for the Chinese group: 44–53%). Compressing the OMBD caused a decrease in word identification by about 20% both for the 20-ms and the 40-ms OMBD. When the OMBD was stretched by a factor of 4 or 8, mosaic speech was basically unintelligible.

The first purpose of this experiment was to measure the intelligibility of mosaic speech in which OMBD was either compressed, preserved, or stretched in time. The results of the three language groups were broadly similar as the results in the preliminary experiment. That is, the intelligibility of English mosaic words was at the highest level at 20- or 40-ms MBD when the OMBDs were preserved or stretched. A similar results also were found in the previous study with Japanese mosaic speech sentences in which the intelligibility was near-perfect for segment durations of up to 40 ms (Nakajima et al., 2018). Kojima and Nakajima's (2016) study also found that the intelligibility of English mosaic speech sentences was rather high (about 80%) at 20-ms or 40-ms temporal segment duration with native-Japanese as listeners.

The second purpose of this experiment was to investigate whether the effects of compressing, preserving, or stretching the mosaic speech would be similar among listeners with different language backgrounds. As expected, mosaic speech intelligibility was significantly higher for the native-English group than for the two non-native English-speaking

(Indonesian and Chinese) groups in each OMBD. Nevertheless, compressing or stretching the speech resulted in similar trends in intelligibility among the three language groups. The intelligibility was relatively high for the preserved OMBDs of 20 and 40 ms for all language groups, and also for the stimuli with the 20-ms OMBD stretched into the MBD of 40 ms, but for the native-English group only. In conclusion, even when the same acoustic information was given to the listeners, intelligibility decreased when the speed of the speech was changed compared to when it was preserved. However, the intelligibility did not change significantly even when the amount of information must have changed (the OMBD of 20 and 40 ms contained different amounts of information) for the English and the Indonesian group at the MBD of 40, 80, or 160 ms. This is the same tendency as found in the preliminary experiment.

Regarding the intelligibility comparison between the language groups, the intelligibility was almost similar in between the non native-English listeners (the Indonesian and the Chinese group), while, as expected, the native-English group showed the highest intelligibility in any preserved/stretched MBD for both OMBDs, except at an MBD of 160 ms for stimuli with a 20-ms OMBD. Regardless of language background, linguistic information in English is thus conveyed relatively well when presented within temporal blocks of 20 and 40 ms, either at a preserved speed or a stretched speed.

In addition, regarding the intelligibility of phonemes of each English word in two types of mosaicizing phase, i.e., the half-phase and the whole-phase, in general some issues for further research were found. One interesting result was that some of the phonemes, i.e., /s/, /r/, /j/, /m/, and /n/, were still highly intelligible even for longer MBD durations, that was, when the OMBD of 20 ms was stretched into 160 ms for the native-English group and the Chinese group (for the phoneme /s/ only). Meanwhile, the phonemes /b/, /h/, and /l/ were difficult to identify for both OMBDs for all language groups, and the phonemes of the stop consonants were more confusing to distinguish from each other, specially for the Indonesian and the

Chinese groups. However, overall, the phoneme intelligibility of initial consonants was less influenced by the MBD, but more influenced by the following phonemes.

Chapter 4 - General Discussion and Conclusions

This present thesis research investigated temporal aspects of speech processing by performing intelligibility tests of mosaiced English words, in which the OMBD was compressed, preserved, or stretched in time. These manipulations did not change the acoustic information but changed the speed of the speech. Two experiments were conducted, a preliminary experiment (Chapter 2) and a main experiment (Chapter 3), with listeners with three different language backgrounds: native-English, Indonesian (both in the preliminary and the main experiment), and Chinese (only in the main experiment). The listeners typed what they had heard using the English alphabet.

The results showed that, overall, for each language group in both experiments, the highest intelligibility was obtained when the OMBD of the stimuli was preserved/stretched into 20 or 40-ms MBDs. The intelligibility decreased as the OMBDs were changed, either compressed or stretched. The most important finding was that if the MBD for presentation was not longer than 40 ms, the intelligibility of mosaic speech was at the highest level when the OMBD was preserved or stretched. This finding similar with the results of several studies during the past decade. The preliminary study of Kojima and Nakajima (2016), Kojima et al. (2017) with English mosaic speech showed that the intelligibility was high when the temporal segment duration was 20 or 40 ms. Their studies used English mosaic speech sentences, which were presented to native-Japanese listeners. In the study of Nakajima et al., (2018), the intelligibility of Japanese mosaic speech sentences was near-perfect when the temporal segment duration was 40 ms or shorter. A study on “pixelated speech” also reported a comparable result, in that the intelligibility of German speech sentences was almost similar to the original speech intelligibility when the segment duration was 50 ms or shorter (Schlittenlacher et al., 2019). Other studies related to this present thesis research investigated

locally time-reversed speech sentences in German (Steffen & Werrani, 1994), English (Saberi & Perrot, 1999), or Japanese (Ueda et al., 2017; Nakajima et al., 2018). In all these studies, the intelligibility became very high for segments of about 40 ms or shorter for stimuli with a normalized speech rate.

The results of the two experiments described in this thesis agreed with Nakajima et al.'s (2018) argument about the resemblance between the block duration of mosaic speech and the frame duration of motion pictures. The segment duration of 40 ms is similar as that employed in movies, in which visual motion is induced by presenting successive static pictures at the same intervals of 24 frames per second (i.e., frames of 42 ms, Read & Meyer, 2000). They suggested that the temporal resolution of about 40 ms is necessary to perceive motion in general. Furthermore, as mentioned in the Introduction (Section 1.6), this temporal segment size seems to be compatible with neural oscillations of 20–33 ms, which are considered to be involved in preserving phonemic intelligibility (Giraud & Poeppel, 2012; Chait et al., 2015). Given the potential correspondence between the present intelligibility data and neuroscientific findings, it is feasible that mosaic speech can be used as an alternative to sounds in current hearing tests. Most testing nowadays is performed with pure tones or natural speech audiometry. With mosaic speech as test stimulus, it will be possible to more systematically assess how temporal and spectral aspects of speech processing develop or change over age, along with possible changes in human cortical functioning and vitality.

The second purpose of this thesis research was to investigate whether the effects of compressing, preserving, or stretching mosaic speech varies among listeners with different language backgrounds. The results of the preliminary experiment were not included in this comparison because there were differences in English words used as stimuli and also the methods. In general, mosaic speech intelligibility was significantly higher for the native-English group than for the non-native English groups (Chapter 3) in each OMBD. However,

by compressing or stretching the OMBD, similar trends in intelligibility appeared in the three language groups in the main experiment; also for the Indonesian group in the preliminary experiment (Chapter 2). That is, the intelligibility was relatively high for the preserved OMBDs of 20 and 40 ms for all language groups, and also for the stimuli with an MBD of 40 ms after stretching the OMBD of 20 ms, but only for the native-English group in the main experiment. As a conclusion, the intelligibility decreased as the speed of the speech was changed by compressing or stretching the OMBD, even though the acoustic information from the preserved OMBD was not changed by these manipulations.

Furthermore, regarding the speed of speech, it was mentioned earlier that by compressing or stretching the OMBD, the mosaic speech would become faster or slower compared to the preserved OMBD. Both experiments thus showed that the non-native English groups preferred the preserved speed of the words, in that the intelligibility of preserved mosaic speech was highest. By contrast, in a different speech context, several studies showed that most people prefer to hear slower speech than the average speech speed, especially when in certain degraded listening conditions, such as under noise or reverberation (Beasley et al., 1972; Konkle et al., 1977; Schmitt & McCroskey, 1981; Schmitt, 1983; Wingfield & Ducharme, 1999; Moore et al., 2007). The performance of the native-English group seems to agree with this, showing that when the speed of speech was slightly slower than the preserved speed, the intelligibility increased, but did not differ significantly from the intelligibility of the preserved OMBD. This happened when the OMBD of 20 ms was stretched into 40 ms.

For people with a hearing impairment, speech intelligibility decreases with increasing speed of speech (Luterman et al., 1966; Sticht & Gray, 1969; Gordon-Salant & Fitzgibbons, 2001; Versfeld & Dreschler, 2002). Also for listeners with normal hearing with mosaic speech, which is a kind of degraded speech, intelligibility decreased when the speech became faster. By mosaicizing, the intelligibility of the speech is degraded as its temporal resolution is

degraded. The speed of speech after mosaicing changes according to its OMBD—speech with an OMBD of 40 ms is slightly slower than speech with an OMBD of 20 ms. Therefore, when the OMBD was changed, either compressed or stretched, the intelligibility changes also and becomes lower. In conclusion, the results of the native-English and the non-native groups were similar when the mosaic speech was speeded-up, preserved, or slowed-down by compressing, preserving, or stretching the OMBDs; except at the OMBD of 20 ms, for which was the native-English group retained their hearing acuity.

Furthermore, regarding the stimuli with the same MBs among both OMBDs, the results showed that the intelligibility did not change significantly for stimuli with the same MBD of 40, 80, or 160 ms, even though the two OMBDs contained different amounts of information. This was found for the native-English and the Indonesian groups in the main experiment (Chapter 3) and also for the Indonesian group in the preliminary experiment (Chapter 2). Thus, these findings suggest that the intelligibility was not affected by OMBD when it was preserved/stretched in the range of 40-160 ms, except for the Chinese group.

In general, manipulating the OMBDs, either by compressing or stretching, showed a similar trend in the intelligibility of English mosaic speech among the three language groups in both experiments, that is, the intelligibility became lower compared to preserved mosaic speech. However, for the native-English group and the Indonesian group (in the Preliminary experiment), although the intelligibility was lower, it was not significantly different from that of the preserved OMBD when the OMBD of 20 ms was stretched into 40 ms. Overall, even though the linguistic information was preserved, intelligibility decreased when the MBD was 80 ms or longer.

Before considering practical implications of mosaic speech, the following needs to be addressed first. In the previous study with Japanese mosaic speech (Nakajima et al., 2018), the intelligibility was near-perfect (> 95%) for native-Japanese listeners. However, in the present

thesis research with English mosaic speech, the native-English group reached intelligibility scores of only about 75% on average. It is possible that the complexity of the English speech sounds, as discussed in the Introduction (see Section 1.3), also affected intelligibility. Therefore, it would be interesting to further investigate this issue by generating mosaic speech in various other languages too, in order to see how intelligibility of mosaic speech varies with language type. It also would be possible to address the intelligibility of English mosaic speech by comparing our present data with those from objective, automated speech recognition systems (Fontan et al., 2017).

Furthermore, a second reason for the relatively low intelligibility scores of the native-English group needs to be addressed too. Different from the previous study with English mosaic speech (Kojima & Nakajima, 2016; Kojima et al., 2017) and Japanese mosaic speech (Kojima et al., 2017; Nakajima et al., 2018), the present thesis research used words and not sentences. The word's intelligibility depends on the intelligibility of its phoneme as suggested in this thesis research (Section 3.3.5), although more research is necessary with statistical tests. This agreed that identifying a word in isolation is not as easy as identifying a word in a sentence context (Marslen-Wilson, 1984). A word in a sentence can be identified from the English syntactic structure (Miller & Isard, 1963) and the semantic context in congruent sentences (Kalikow et al., 1977), which can assist the word identification quickly compared to the word-alone presentation (Miller et al., 1951; Grosjean, 1980; Salasoo & Pisoni, 1985).

Overall, this present thesis research was conducted to investigate what temporal resolution is required for speech perception. It is expected that the findings of this thesis research will be useful for research relate speech and resource for researchers.

In conclusion, the present research showed that:

- The intelligibility of English mosaic speech was relatively high for the preserved OMBDs of 20 ms and 40 ms for all language groups. The intelligibility of English mosaic speech

decreased when the speed of the speech was changed in comparison to when it was preserved, even though the same acoustic information was given to the listeners. There was one exception for the native-English and the Indonesian group (in the Preliminary experiment). In these groups, the intelligibility was preserved when the OMBD of 20 ms was stretched into 40 ms. Thus, this thesis research suggests that presenting the same acoustic information in any temporal segment does not guarantee that the intelligibility will be preserved, but the temporal segment duration plays the most important role to determine intelligibility in mosaic speech perception. In other words, the intelligibility was affected by mosaic block duration (MBD).

- The intelligibility of English mosaic speech did not change significantly even when the amount of information must have changed (the OMBD of 20 ms and 40 ms contained different amounts of information) at the same MBD of 40, 80, or 160 ms for the native-English and the Indonesian listeners, but not for the Chinese listeners. Thus, this thesis research suggests that intelligibility was not affected by OMBD when it was preserved/stretched in the range of 40-160 ms.
- Based on the findings with mosaic speech, this thesis research suggests that humans can extract linguistic information from individual speech segments of about 40 ms, but that there is a limit to the amount of linguistic information that can be conveyed within a block of about 40 ms or below.

References

- Abdi, H. (2010). Holm's Sequential Bonferroni Procedure. In Salkind, N. J. (Editor). *Encyclopedia of Research Design* (pp. 573–577). Sage: Thousand Oaks, California, United States of America.
- Anderson, S., Skoe, E., Chandrasekaran, B., & Kraus, N. (2010). Neural timing is linked to speech perception in noise. *Journal of Neuroscience*, *30*(14):4922-6. doi: 10.1523/JNEUROSCI.0107-10.2010.
- Apoux, F., & Bacon, S. P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *The Journal of the Acoustical Society of America*, *116*(3):1671-80.
- Baker, S., Weinrich, B., Bevington, M., Schroth, K., & Schroeder, E. (2008). The effect of task type on fundamental frequency in children. *International Journal of Pediatric Otorhinolaryngology*, *72*, 885–889.
- Benard, M. R., & Başkent, D. (2013). Perceptual learning of interrupted speech. *PLoS One*, *8*(3):e58149. doi: 10.1371/journal.pone.0058149.
- Bogardus, S. T. Jr., Yueh, B., & Shekelle, P.G. (2003). Screening and management of adult hearing loss in primary care: clinical applications. *The Journal of American Medical Association*, *289*(15):1986-90.
- Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking clearly for children with learning disabilities: sentence perception in noise. *The Journal of Speech, Language, and Hearing Research*, *46*(1):80-97.
- Broersma, M., & Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise. *Speech Communication*, *52*, Issues 11–12, pp. 980-995. doi:10.1016/j.specom.2010.08.010
- Cambridge Dictionary. Available online: <https://dictionary.cambridge.org/> (accessed on 26–29 July 2019).
- Carley, P., Mees, I. M., Collins, B. Basic concepts. (2018). In *English Phonetics and Pronunciation Practice*. Routledge: London, United Kingdom; New York, United States of America, pp. 1–2.
- Carre, R., Divenyi, P., & Mrayati, M. (2017). *Speech: A dynamic process*. Walter de Gruyter: Berlin, Germany.
- Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdiscip Reviews: Cognitive Science*, *1*(5): 629–647. doi:10.1002/wcs.63.

- Chait, M., Greenberg, S., Arai, T., Simon, J.Z., & Poeppel, D. (2015). Multi-time resolution analysis of speech: Evidence from psychophysics. *Frontiers in Neuroscience*, 9, 214, doi:10.3389/fnins.2015.00214.
- Chermak, G. D., & Musiek, F. E. (1997). Central auditory processing disorders: New perspectives, 1st edition. Singular Publishing Group, Inc.: San Diego, California, United States of America.
- Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd edition). Academic Press: New York, United States of America.
- Cohen, J. (1992). A power primer. *Psychological Bulletin* 112(1), 155–159. doi:10.1037/0033-2909.112.1.155.
- Colton, R. H., & Casper, J. K. (1990). Understanding voice problems: a physiological perspective for diagnosis and treatment (3rd edition). Lippincott Williams and Wilkins: Baltimore, United States of America.
- Crespo, J., & Henriks, R. (2014). Speech reinforcement in noisy reverberant environments using a perceptual distortion measure. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 910–914.
- Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1011-21.
- Denes, P. B., & Pinson, E. N. (2015). The speech chain, the physics and biology of spoken language, 2nd Edition. Waveland Press, Inc.: Illinois, United States of America.
- De Saussure, F. (1966). Course in general linguistics. McGraw-Hill Book Company: New York, United States of America.
- Diehl, R. L. (2008). Acoustic and auditory phonetics: The adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society B*, 363(1493):965-78. doi:10.1098/rstb.2007.2153
- Ding, N., Patel, A.D., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, 81, 181–187.
- Dirks, D. D., Wilson R. H., & Bower, D.R. (1969). Effect of pulsed masking on selected speech materials. *The Journal of the Acoustical Society of America*, 46(4):898-906.
- Dong, H., & Lee, C. (2018). Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering. *EURASIP Journal on Audio, Speech, and Music Processing*, 3, 1–13.

- Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, *102*(4):2403-11. doi: 10.1121/1.419603.
- Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, *95*, 2670–2680. doi:10.1121/1.408467.
- Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, *95*, 1053–1064. doi:10.1121/1.
- Du, A., Lin, C., & Wang, J. (2015). Effect of speech rate for sentences on speech intelligibility. *2014 IEEE International Conference on Communication Problem-Solving*, 233-236. 10.1109/ICCPS.2014.7062261.
- Ellermeier, W., Kattner, F., Ueda, K., Doumoto, K., & Nakajima, Y. Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands. *The Journal of the Acoustical Society of America*, *138*, 1561–1569. doi:10.1121/1.4928954.
- Fairbanks, G., & Kodman, F. Jr. (1957). Word intelligibility as a function of time compression. *The Journal of the Acoustical Society of America*, *29*, 636–641.
- Fant, G. (1973). *Speech sounds and features*. MIT Press: Cambridge, Massachusetts, United States of America.
- Fastl, H., & Zwicker, E. (2007). Critical Bands and Excitation. In *Psychoacoustics: Facts and Models*, 3rd edition (pp. 149–172). Springer: New York, United States of America.
- Feldman, H. M. (2019). How young children learn language and speech: Implications of theory and evidence for clinical pediatric practice. *Pediatrics in Review*, *40*(8): 398–411. doi:10.1542/pir.2017-0325.
- Field, A. (2009). Non-parametric Tests. In *Discovering Statistics Using SPSS*, 3rd edition, Sage Publication: London, United Kingdom.
- Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *131* (2), pp. 1490–1501.
- Fontan, L., Ferrané, I., Farinas, J., Piquier, L., Tardieu, J., Magnen, C., Gaillard, X. A., & Füllgrabe, C. (2017). Speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research*, *60*, 2394–2405.

- Gallun, F., & Souza, P. (2008). Exploring the role of the modulation spectrum in phoneme recognition. *Ear & Hearing, 29*(5):800-13.
- Garrett, M. F. (1978). Word and Sentence Perception. In *Perception* (pp 611-625). Springer-Verlag: Berlin Heidelberg, Germany.
- Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*, 511–517, doi:10.1038/nn.3063.
- Greenberg, S., Arai, T. (2004). What are the essential cues for understanding spoken language? *IEICE Transactions on Information and Systems, E87-D*, 1059–1070, doi:10.1121/1.4744396.
- Grosjean, F. (1980). Spoken word recognition processes and gating paradigm. *Perception & Psychophysics, 28*, 267–283. doi:10.3758/BF03204386
- Harmon, L. D. (1973). The recognition of faces. *Scientific American, 229*, 71–82.
- Harris, R. W., & Swenson, D. W. (1990). Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment. *Audiology, 29*(6):314-21. DOI: 10.3109/00206099009072862
- Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in System Neuroscience, 8*(35). doi: 10.3389/fnsys.2014.00035.
- Hirsch, H. G. (1992). Robust speech recognition in noisy and reverberant environments. In *Speech Recognition and Understanding. NATO ASI Series, Vol. 75*. pp. 101-106. Springer-Verlag: Heidelberg, Berlin, Germany.
- Hochmair-Desoyer, I. J., Hochmair, E. S., Fischer, R. E. & Burian, K. (1980). Cochlear prostheses in use: recent speech comprehension results. *The European Archives of Oto-Rhino-Laryngology, 229*, pp. 81-98.
- Hochmair-Desoyer, I. J., Hochmair, E. S. & Stiglbanner, H. K. (1985). Psychoacoustic temporal processing and speech understanding in cochlear implant patients. In *Cochlear Implants* (editors R. A. Schindler & M. M. Merzenich), pp. 291-304. Raven Press: New York, United States of America.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.
- Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention Perception & Psychophysics, 72*(5): 1218–1227. doi:10.3758/APP.72.5.1218.

- Hood, J. D. (1950). Studies in auditory fatigue and adaptation. *Acta oto-laryngologica, Supplementum*, 92:1–57
- Jeddi, Z., Lotfi, Y., Moossavi, A., Bakhshi, E., & Hashemi, S. B. (2019). Correlation between Auditory Spectral Resolution and Speech Perception in Children with Cochlear Implants. *Iranian Journal of Medical Sciences*, 44(5):382-389. doi: 10.30476/IJMS.2019.44967.
- Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61,1337–1351.doi:10.1121/1.381436.
- Kellogg, E. W. (1939). Reversed speech. *The Journal of the Acoustical Society of America*, 10, 324–326.
- Kidd, G. R, & Humes, L.E. (2012). Effects of age and hearing loss on the recognition of interrupted words in isolation and in sentences. *The Journal of the Acoustical Society of America*, 131(2):1434-48.
- Killion, M. C. (1997). Hearing aids: past, present, future: moving toward normal conversations in noise. *British Journal of Audiology*, 31(3):141-8.
- Kishida, T., Nakajima, Y., Ueda, K., & Remijn, G. (2016). Three factors are critical in order to synthesize intelligible noise-vocoded Japanese speech. *Frontiers in Psychology*, 7, 517, doi:10.3389/fpsyg.2016.00517
- Kojima, K., & Nakajima, Y. (2016). Influence of the temporal-unit duration on the intelligibility of English speech. In *The 31st International Congress of Psychology (ICP), Volume 51, Issue S1*, <https://doi.org/10.1002/ijop.12378>.
- Kojima, K., Nakajima, Y., Ueda, K., Remijn, G. B. Elliott, M. A., & Arndt, S. (2017). Influence of the temporal-unit duration on the intelligibility of mosaic speech: A comparison between Japanese and English. In *Proceedings of the 33rd Annual Meeting of the International Society for Psychophysics*, Fechner Day 2017, p. 127.
- Kress, J. E., & Fry, E. B. (2016). *The reading teacher's, book of list*, 6th edition. Jossey-Bass: San Francisco, United States of America, pp. 21–171.
- Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14, 677- 709.
- Laures, J., & Weismer, G. (1999). The effect of flattened F₀ on intelligibility at the sentence-level. *The Journal of Speech, Language, and Hearing Research*, 42, 1148–1156.
- Leddy, M. The biological bases of speech in people with Down syndrome. In Miller J, Leddy M, & Leavitt LA. (Editors). *Improving the Communication of People with Down*

- Syndrome. Paul H Brookes Publishing: Baltimore, Maryland, United States of America 1999. pp. 61–80.
- Liberman, A. M. (1957). Some Results of Research on Speech Perception. *The Journal of the Acoustical Society of America*, Volume 29, Number 1.
- Liberman, A. M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431-61.
- Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1–36, doi:10.1016/0010-0277(85)90021-6.
- Marslen-Wilson, W. D. (1984). Function and process in spoken word-recognition. In Bouma, H., Bouwhuis, G. (Editors). *Attention and Performance X: Control of Language Processes* (pp. 125–150). Erlbaum: Hillsdale, New Jersey, United States of America.
- Mei, X. D., & Sun, S. H. (2001). Silence and speech segmentation for noisy speech using a wavelet based algorithm. *Chinese Journal of Electronics* 10(4):439-443.
- Meyer-Eppler, W. (1950). Reversed speech and repetition systems as means of phonetic research. *The Journal of the Acoustical Society of America*, 22, 804–806.
- Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 22, 167–173.
- Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41, 329–335. doi: 10.1037/h0062491.
- Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2, 217–228. doi:10.1016/S0022-5371(63)80087-0.
- Moore B. C. J. (1998). *Cochlear hearing loss*. Whurr Publishers: London, United Kingdom.
- Moore, B. C. J. (2008). Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):947-63. doi: 10.1098/rstb.2007.2152.
- Moore, B. C. J., Tyler, L. K., & Marslen-Wilson, W. D. (2007). *The perception of speech: from sound to meaning*. Oxford University Press: New York, United States of America.
- Moradi, S., Lidestam, B., Saremi, A. & Rönnberg, J. (2014). Gated auditory speech perception: effects of listening conditions and cognitive capacity. *Frontiers in Psychology*, 5(531). <https://doi.org/10.3389/fpsyg.2014.00531>.

- Nabelek, A. K. (1993). Communication in noisy and reverberant environments. In Stuebelaker GA., & Hochberg I. (Editors). *Acoustical Factors Affecting Hearing Aid Performance*. Needham Heights, Allyn and Bacon: Boston, Massachusetts, United States of America, pp. 15–28.
- Nakajima, Y., Matsuda, M., Ueda, K., & Remijn, G. B. (2018). Temporal resolution needed for auditory communication: Measurement with mosaic speech. *Frontiers in Human Neuroscience*, *12*, 149, doi:10.3389/fnhum.2018.00149.
- Neuman A, & Hochberg, I. (1983). Children's perception of speech in reverberation. *The Journal of the Acoustical Society of America*, *73*:215–2149.
- Nusbaum, H. C., & Magnuson, J. (1997). Talker normalization: phonetic constancy as a cognitive process. In K. Johnson & J. W. (Editors). *Mullennix Talker Variability in Speech Processing*. Academic Press: San Diego, California, United States of America, 109–129.
- Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Editors). *Pattern Recognition by Humans and Machines, Speech Perception (Volume 1)*. Academic Press: San Diego, California, United States of America, 113–157.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3):355-76. doi: 10.3758/bf03206860.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific perceptual learning in spoken word recognition. *Perception & Psychophysics*, *60*, 355–376. doi: 10.1121/1.397688.
- Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech Communication*, *4*(1-3): 75–95. doi: 10.1016/0167-6393(85)90037-8.
- Prendergast, G., Johnson, S. R., & Green, G. G. R. (2008). Amplitude modulation shape and speech intelligibility. *The Journal of the Acoustical Society of America* *123*(5):3734. doi: 10.1121/1.2935237
- Read, P., & Meyer, M. P. (2000). Cinematographic Technology. In *Restoration of Motion Picture Film*, 1st edition (pp. 9–45). Butterworth-Heinemann Elsevier Ltd.: Oxford, United Kingdom.
- Richards, J.C., & Schmidt, R.W. (2010) *Longman dictionary of language teaching & applied linguistics*, 4th edition. Routledge: Abingdon, London, United Kingdom; New York, United States of America, p. 126.

- Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 336(1278):367-73.
- Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research* (2nd edition). Sage Publication: Newbury Park, California, United States of America.
- Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, 398, 760.
- Salasoo, A., & Pisoni, D. (1985). Interaction of knowledge source in spoken word identification. *Journal of Memory and Language*, 24,210–231.doi:10.1016/0749-596X(85)90025-7
- Santi, S., Nakajima, Y., Ueda, K., & Remijn, G. B. (2019). Effects of compressing or stretching mosaic block duration on intelligibility of English mosaic speech. *In Proceedings of the 35th Annual Meeting of the International Society for Psychophysics*, Fechner Day 2019, p. 35.
- Schlittenlacher, J., Staab, K., Çelebi, Ö., Samel, A., & Ellermeier, W. (2019). Determinants of the irrelevant speech effect: Change in spectrum and envelope. *The Journal of the Acoustical Society of America*, 145, 3625–3632. <https://doi.org/10.1121/1.5111749>.
- Schlueter, A., Lemke, U., Kollmeier, B., & Holube, I. (2016). Normal and time-compressed speech: How does learning affect speech recognition thresholds in noise? *Sage Publication*, 20: 1-13. doi: 10.1177/2331216516669889.
- Shafiro, V., Sheft, S., & Risley, R. (2011). Perception of interrupted speech: effects of dual-rate gating on the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, 130(4):2076-87.
- Shafiro, V., Sheft, S., & Risley, R. (2016). The intelligibility of interrupted and temporally altered speech: Effects of context, age, and hearing loss. *The Journal of the Acoustical Society of America*, 139, 455–465, doi:10.1121/1.4939891.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303-4.
- Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303–304.
- Smith, R. H. (2012). New direction in speech perception. In *Bloomsbury Companion to Phonetics*. Editors: Rachael-Anne Knight, Mark J. Jones. Bloomsbury: London, British. pp. 389-412.
- Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416, 87–90, doi:10.1038/416087a.

- Souza, P. E., Boike, K. T., Witherell, K., Tremblay, K. (2007). Prediction of speech recognition from audibility in older listeners with hearing loss: effects of age, amplification, and background noise. *The Journal of the American Academy of Audiology*, 18(1):54-65. 2007;18(1):54-65.
- Steffen, A., & Werani, A. (1994). Ein experiment zur zeitverarbeitung bei der sprachwahrnehmung. In Kegel, G., Arnhold, T., Dahlmeier, K., Schmid, G., & Tischer, B. (Editors). *Sprechwissenschaft & Psycholinguistik* (Volume 6, pp. 189–205). Westdeutscher Verlag: Opladen, Germany.
- Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111, 1872–1891, doi:10.1121/1.1458026.
- Strait, D. L., Parbery-Clark, A., & O'Connell, S., & Kraus, N. (2013). Biological impact of preschool music classes on processing speech in noise. *Developmental Cognitive Neuroscience*, 6:51-60.
- Ueda, K., Nakajima, Y., Ellermeier, W., & Kattner, F. (2017). Intelligibility of locally time-reversed speech: A multilingual comparison. *Scientific Reports*, 7, 1782, doi:10.1038/s41598-017-01831-z.
- Volín, J., & Skarnitzl, R. (2018). Foreign Accents and English in International Contexts. In the Pronunciation of English by Speakers of Other Languages; Cambridge Scholars Publishing: Newcastle Upon Thyne, United Kingdom, 2018; pp. 1–2.
- Waibel, A. (1987). Prosodic knowledge sources for word hypothesization in a continuous speech recognition system. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*, pp. 856–859.
- Warrier, C. M., Johnson, K. L., Hayes, E. A., Nicol, T., & Kraus, N. (2004). Learning impaired children exhibit timing deficits and training-related improvements in auditory cortical responses to speech in noise. *Experimental Brain Research*, 157(4):431-41. doi: 10.1007/s00221-004-1857-6.
- Wells, J. C. (1982). *Accent of English*; Cambridge University Press: Cambridge, United Kingdom.
- Wells, J. C. (2014). *Longman Dictionary of Contemporary English*, 6th edition. Pearson: London, United Kingdom.
- Wells, J. C. (2008). *Longman Pronunciation Dictionary*, 3rd edition. Pearson: London, United Kingdom.

- Wenanda, D., Suryani, S. (2016). Analisis Kesalahan Berbahasa Inggris pada Tataran Fonologis. In Prosodi: *Jurnal Ilmu Bahasa dan Sastra, Volume X* (Nomor 2, pp. 145–155). Department of English, University of Trunojoyo: Madura, Indonesia.
- Whalen, D. H. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception and Psychophysics* 46(3), 284-292. doi: 10.3758/bf03208093.
- Wingfield, A. (1996). Cognitive factors in auditory performance: context, speed of processing, and constraints of memory. *Journal of the American Academy of Audiology*, 7:175–182.
- Wingfield, A., & Ducharme, J. L. (1999). Effects of age and passage difficulty on listening-rate preferences for time-altered speech. *The Journals of Gerontology*, 54:199–202.
- Wingfield, A., McCoy, S., Pelle, J., Tun, P., & Cox, L. (2006). Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity. *Journal of the American Academy of Audiology*, 17:487–497.
- Wingfield, A., Tun, P., Koh, C., & Rosen, M. (1999). Regaining lost time: adult aging and the effect of time restoration on recall of timecompressed speech. *Psychology and Aging*, 14:380–389.
- Wróblewski, M., Lewis, D., Valente, D. L., & Stelmachowicz, P. G. (2012). Effects of Reverberation on Speech Recognition in Stationary and Modulated Noise by School-Aged Children and Young Adults. *Ear & Hearing*, 33, 731–744
- Yoo, S., Boston, J., El-Jaroudi, A., & Li, C. (2007). Speech signal modification to increase intelligibility in noisy environments. *The Journal of the Acoustical Society of America*, 122, 1138–1149.

Appendix A. Informed consent and instruction of stimulus recording (preliminary experiment)

Dear Participant,

Thank you for your participation in English Mosaic Speech study.

The topic of my study is “**Intelligibility of English Mosaic Speech**”

We are going to conduct a research but before that, we have to create speech sounds (speech recording). The original of speech (WAV file) will be transformed into mosaic speech and then will become stimulus in the research experiment.

We need some information about you and we guarantee your privacy.

Name :
Born Place :
Age :
Education :

Nationality

Father :

Mother :

Mother Tongue:

Dialect:

Please write down all the languages you used in your daily life and its fluency level!

Answer:

Experiment instructions:

1. Before recording, I will measure the distance between recorder and your mouth.
2. Please practice reading the list of words first to avoid mistakes during recording.
3. Please read the words in word by word with a space between the words about 2-5 seconds and each word will be recorded 3 times.
4. Please read the word about 2 seconds after the record button is pressed for each group word. There will be about 10 seconds as a space between the groups.
5. If you make a mistake, please say “mistake” and read the same word again.

Santi

Nakajima Laboratory

Department of Human Science

Graduated School of Design, Kyushu University

Tel.: 070-4750-1710

Email: santi.dp17@gmail.com

Appendix B. Stimulus recording procedure

1. Preliminary experiment



2. Main experiment



Appendix C. Informed consent and instruction of listening experiment (Preliminary experiment)

Dear Participant,

Thank you for your participation in English Speech study.

The topic of my study is “**Intelligibility of English Speech**”

We are going to conduct a listening experiment. I will present some sound stimuli. The sound stimuli are degraded English single words.

We need some information about you and we guarantee your privacy.

Name :

City :

Age :

Education :

Nationality

Father :

Mother :

Mother Tongue:

First Language :

What kind of English language do you speak?

American British Australian Other ()

Please write down all the language(s) (including your mother tongue) you speak frequently (in frequency order)!

Answer:

Do you have normal hearing? [Y] [N]

Santi

Nakajima Laboratory

Department of Human Science

Graduated School of Design, Kyushu University

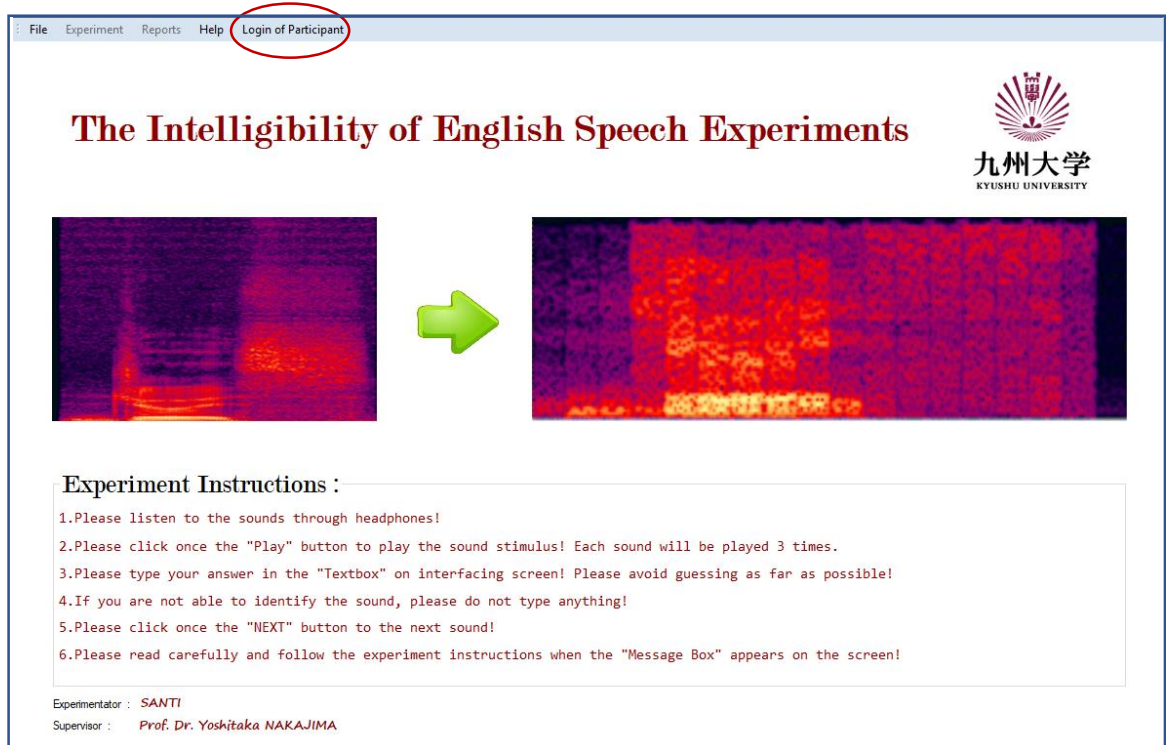
Tel.: 070-4750-1710 / Email: santi.dp17@gmail.com

Experiment instructions:

1. Please listen to the sounds through headphones!
2. Please click once the “PLAY” button to play the sound stimulus! Each sound will be played 3 times.
3. Please type your answer in the "Textbox" on interfacing screen! Please avoid guessing as far as possible!
4. If you are not able to identify the sound, please do not type anything!
5. Please click once the “NEXT” button to the next sound!
6. Please read carefully and follow the experiment instructions when the "Message Box" appears on the screen!

How to use the program:

1. Please use the headphones!
2. Please click “Login of Participant” on bar Menu!



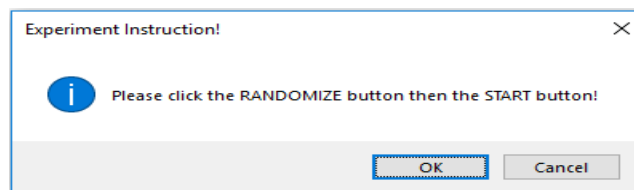
3. Please type your participants' number and ID on the textboxes, then click the “NEXT” button!

Login

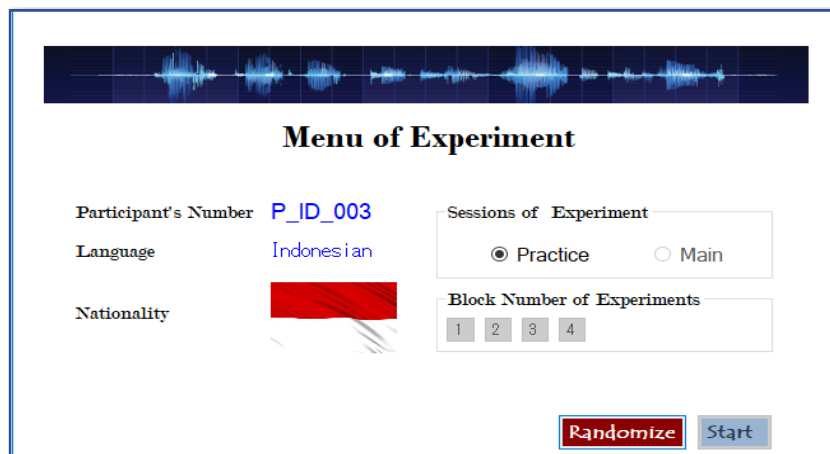
Username
(Please type your participant's number!)

Password
(Please type your Participant's ID!)

4. Please click the “OK” button!



5. Please click once the “RANDOMIZE” button and then the “START” button for starting the practice session!



6. Please click once the “PLAY” button to play the sound stimulus! Each sound will be presented 3 times.

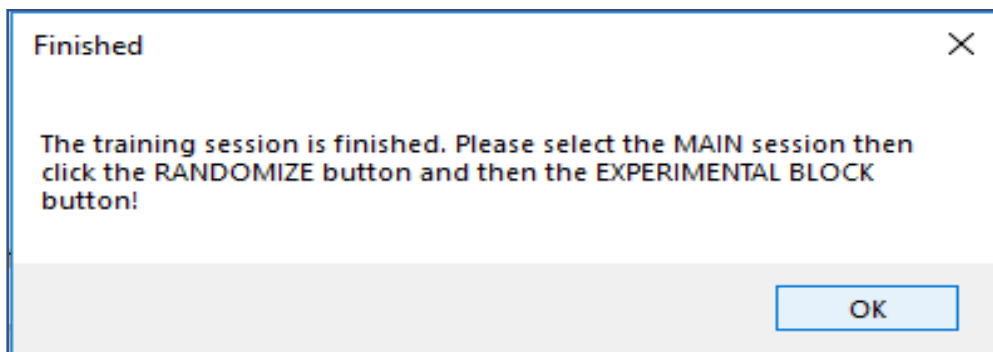


7. After the sound ends, the cursor (pointer) is going to the textbox automatically. Please type what you hear through the headphones! Please avoid guessing answer and if you are not able to identify the sound, please do not type anything!

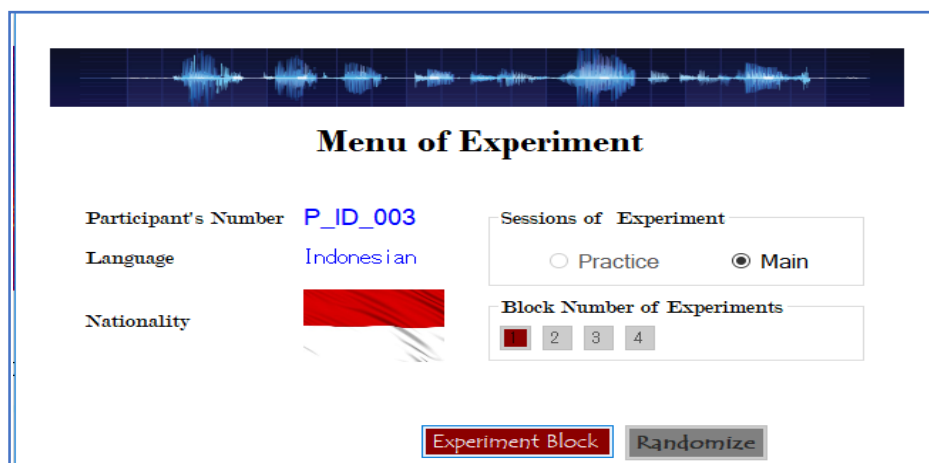
8. Please click once the “NEXT” button to the next sound!

9. Please do the same instructions (5, 6, and 7) until the trial block is finish.

10. After practice session is finished, please click “OK” button to the next session when message box appears!

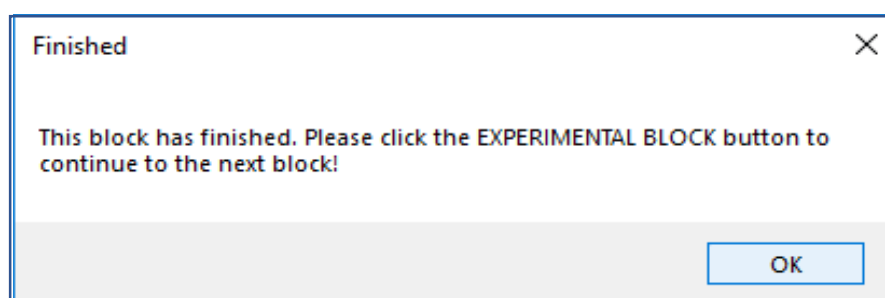


11. Please select the “MAIN” session, then click the “RANDOMIZE” button and then the “EXPERIMENT BLOCK” button.



12. Please do the same instructions (5, 6, and 7) until the trial block is finish.

13. Please click the “OK” button to the next block of experiment when message box appears!



14. Please do the same instructions (10, 11, and 12) until the experiment is finish.

Appendix D. Statistical analysis of preliminary experiment data

1. Normality check of English mosaic speech words intelligibility

	Tests of Normality					
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
CompressOMBD20_HalfPhase	0.227	20	0.008	0.886	20	0.023
CompressOMBD20_WholePhase	0.230	20	0.007	0.826	20	0.002
CompressOMBD40_HalfPhase	0.184	20	0.074	0.912	20	0.068
CompressOMBD40_WholePhase	0.288	20	0.000	0.860	20	0.008
PreserveOMBD20_HalfPhase	0.251	20	0.002	0.826	20	0.002
PreserveOMBD20_WholePhase	0.238	20	0.004	0.880	20	0.018
PreserveOMBD40_HalfPhase	0.192	20	0.053	0.873	20	0.014
PreserveOMBD40_WholePhase	0.232	20	0.006	0.884	20	0.021
Stretch2OMBD20_HalfPhase	0.287	20	0.000	0.863	20	0.009
Stretch2OMBD20_WholePhase	0.233	20	0.006	0.878	20	0.016
Stretch2OMBD40_HalfPhase	0.227	20	0.008	0.886	20	0.023
Stretch2OMBD40_WholePhase	0.255	20	0.001	0.791	20	0.001
Stretch4OMBD20_HalfPhase	0.225	20	0.009	0.866	20	0.010
Stretch4OMBD20_WholePhase	0.214	20	0.017	0.869	20	0.011
Stretch4OMBD40_HalfPhase	0.295	20	0.000	0.667	20	0.000
Stretch4OMBD40_WholePhase	0.255	20	0.001	0.787	20	0.001
Stretch8OMBD20_HalfPhase	0.373	20	0.000	0.622	20	0.000
Stretch8OMBD20_WholePhase	0.291	20	0.000	0.774	20	0.000
Stretch8OMBD40_HalfPhase	0.527	20	0.000	0.351	20	0.000
Stretch8OMBD40_WholePhase	0.538	20	0.000	0.236	20	0.000

a. Lilliefors Significance Correction

2. Friedman Test between all mosaic speech stimulus types

Test Statistics ^a	
N	20
Chi-Square	203.839
df	19
Asymp. Sig.	0.000

a. Friedman Test

**Appendix E. Sound pressure level (SPL) of original speech sounds (Fast peak)
Preliminary experiment)**

No.	Word	Sound pressure level (dBA)	No.	Word	Sound pressure level (dBA)
1	bag	67.8	41	keep	56.3
2	case	63.2	42	nine	66.1
3	fill	64.2	43	raise	63.5
4	red	63.6	44	win	63.2
5	leaf	61.2	45	feed	56.9
6	moon	59.9	46	gate	61.5
7	rate	67.8	47	love	64.6
8	young	70.4	48	map	65.3
9	date	66.8	49	bed	67.4
10	house	70.5	50	cut	65.8
11	king	62.2	51	pick	59.9
12	match	68.6	52	shape	65.6
13	hate	62	53	fat	73
14	kick	60.8	54	june	66.1
15	boy	67.9	55	egg	65.2
16	ride	73.4	56	sing	63.8
17	big	62.1	57	catch	64.5
18	cake	63.1	58	put	63.2
19	heat	58.4	59	rain	64.6
20	wine	70.2	60	seed	60.3
21	hide	68.3	61	fight	69.6
22	lack	67.3	62	name	63.1
23	page	60.5	63	pan	65.1
24	sit	62.5	64	sheep	62.7
25	bus	74.2	65	book	65.3
26	cat	68.8	66	ten	63.2
27	lake	62.6	67	late	66
28	wide	72	68	sad	68.5
29	hit	72.7	69	five	72
30	leg	63.4	70	hat	68.8
31	pain	62.1	71	pull	61.4
32	touch	68.6	72	tape	60.1
33	fan	69.2	73	nice	64.9
34	night	68.1	74	cook	63.2
35	push	65.3	75	fun	69.7
36	shake	64.7	76	wave	67.3
37	eight	61.6	77	fish	63.2
38	lip	62.6	78	line	69
39	mad	64.8	79	pay	62.1
40	run	69	80	seat	60.6

Appendix F. Sound pressure level (SPL) of original speech sounds (Fast peak) Main experiment)

Word	Real speech (dBA)	Original speech sound (Equalized)		Word	Real Speech (dBA)	Original speech sound (Equalized)	
		Before	After			Before	After
bush	74.7	79.7	70.8	push	68.9	74.1	71
love	74.3	79.6	71.6	dive	78.9	84.2	74.2
rage	74	76.9	71.8	juice	67	68.6	69.5
feed	66.4	67.5	67.4	gun	73.9	79.3	71.5
rush	74.5	80.2	73.3	king	65.7	67.8	66.6
date	73.7	76.7	70	touch	69.7	75.2	74
nine	75.7	80.6	73.5	nap	74	78.1	73.6
food	70.2	73.3	67.6	move	66.5	68.8	66.2
size	73.7	79	73.9	name	70.8	73.2	69
yell	76	78.8	72	lab	77	82.8	72.6
fish	70.6	73.9	72	guide	77.5	83.7	73.1
mouse	77.5	82	74.1	chief	68.3	68.9	69.2
tag	73.7	77.9	72.8	youth	70.7	73.4	69.1
beep	70.1	71.9	70.3	hate	68.4	71.1	71
shut	71.5	75.3	74.4	deep	67.1	69	67.3
wing	70.4	72.7	69.2	rise	77.8	83.3	73.6
tooth	68.2	70.5	69	head	66.1	70.4	69.7
doubt	77.9	82	73.7	gum	71.7	77.4	71.2
check	72.4	76.5	73.7	shine	70.5	76	73.5
page	68	70.7	71	choose	73	74.3	70.8
tab	73.7	75.9	71.6	cave	71.9	74.4	70.5
rule	68.7	72.8	67	tell	72.3	76.6	71.3
line	75.5	80.5	73.1	peach	67.5	69.1	69.7
sheep	67.3	68.3	70.3	hat	74.8	79.6	73.8
wife	77.3	82.9	73.6	wish	69.3	73.3	70.4
soup	69.7	70	69.7	cheese	72.2	72.7	70.7
big	70.7	74.1	68.1	game	70.3	72.2	70.1
couch	79.1	83.9	74.7	young	72.2	78	73.1
pig	69.8	73.5	69.3	book	70.6	75.5	70.8
cook	68.5	73.7	71.8	keep	68.1	69	69.7
shape	68.9	71.7	71.1	safe	71.7	74.4	71.6
gel	70.5	74.8	70.7	nice	71.6	76.2	74
rub	76	81.7	72.9	hang	69.7	72.9	71.1
mess	73.7	79.2	73.1	wise	78.2	83.8	73.5
give	70.6	75.3	69.4	loud	77	82.4	73.4
wake	68.3	73	70.8	june	66	68.2	67.5
dish	73.3	76.7	70.8	south	73.5	78.3	74.3
mood	67.1	71.7	66.5	face	73.4	77	72
judge	74.1	79.2	72.7	map	75.2	78.8	72.5
five	77.5	83.5	74	life	75.1	80.5	73.4

Appendix G. Informed consent and instruction of listening experiment (Main experiment)

Participants' number:

Dear Participant,

Thank you for your participation in English Speech study.

The topic of my study is “**Intelligibility of English Speech**”

We are going to conduct a listening experiment. I will present some sound stimuli. The sound stimuli are original and degraded speech of English single words. Before the experiment begins, the experimenter will check your hearing level. The experiment will take approximately 2 hours with 2 batches. Each batch will be started with one trial block section for practicing and the experimenter will be in the room also. Please feel free to ask any question or give comments during and after this practicing block. There will be about 5 minutes for break between the batches. Please follow the experiment instructions and try to get relaxed during the experiment. Please do not try to think hardly and please do not use too much time!

We need some information about you and we guarantee your privacy.

Name :

City/Nationality :

Age/Gender :

Education :

Mother tongue :

Please write down all the languages you used for education and your daily life communication!

Do you have normal hearing? [Yes / No]

How tired are/were you? Please, rate from 0 to 10 (0 = not tired at all, 10 = extremely tired):

1. Before the batch 1 of experiment: _____; after the batch 1 of experiment: _____

2. Before the batch 2 of experiment: _____; after the batch 2 of experiment: _____

Santi

Supervisor: Prof. Yoshitaka Nakajima

Department of Human Science

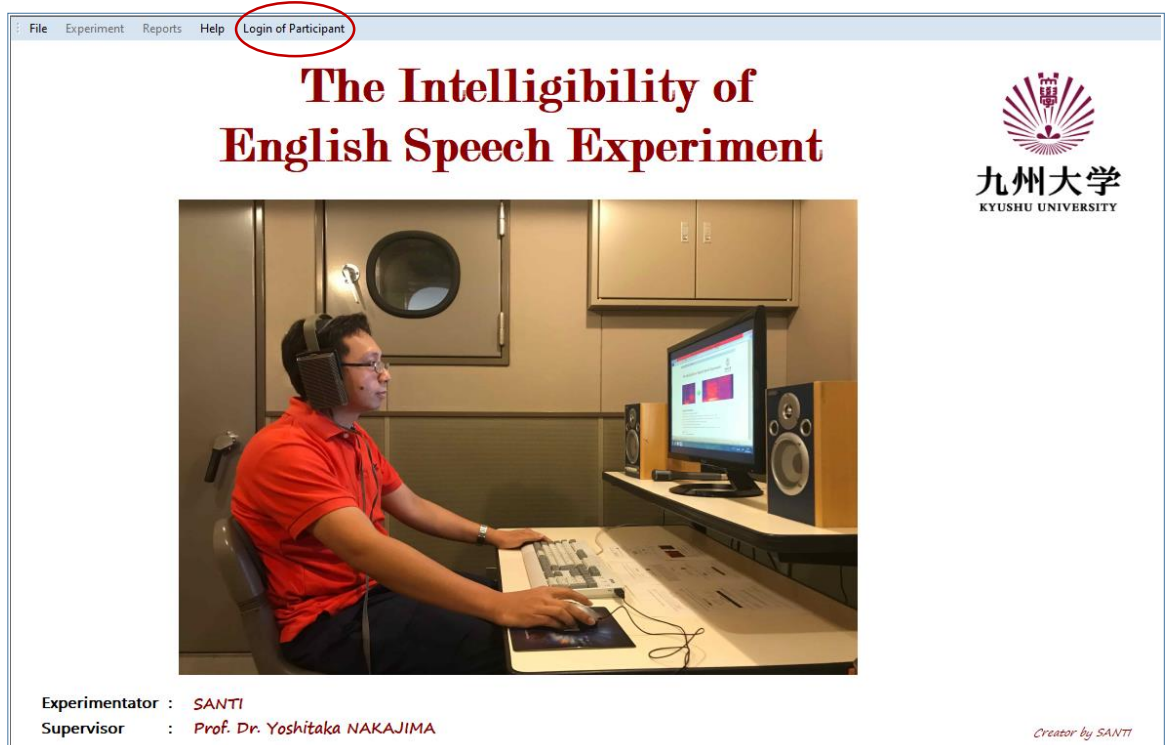
Graduated School of Design, Kyushu University

4-9-1, Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan

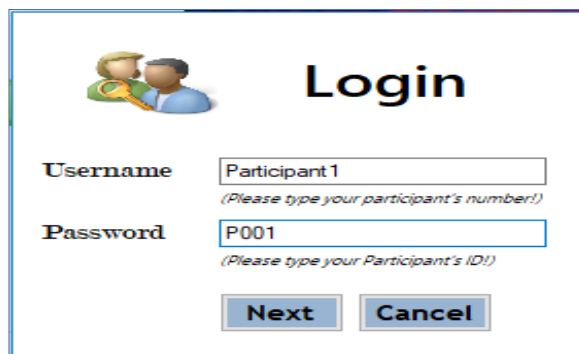
Tel.: +81 70 4750 1710 / Email: santi.dp17@gmail.com

Experiment instructions (how to use the program):

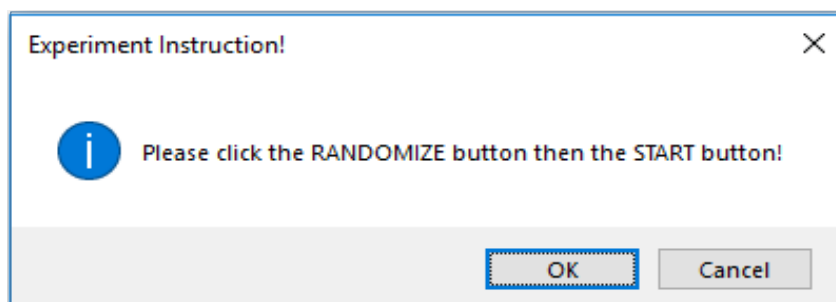
1. Please use the headphones and be sure that it covers your ears tightly and thoroughly.
2. Please click “Login of Participant” on bar Menu!



3. Please type your participant's number and ID on the textboxes, then click the “NEXT” button!



4. Please click the “OK” button to continue!



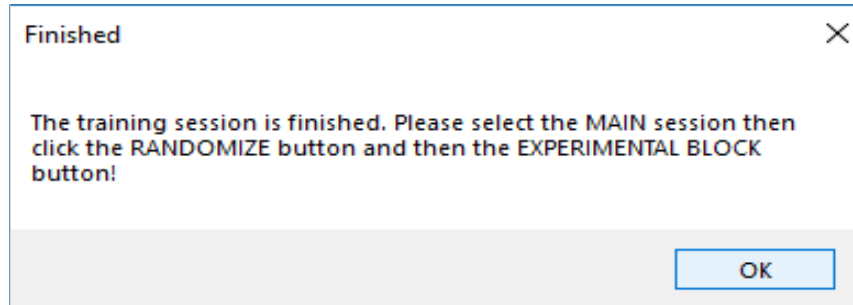
- Please click once the “RANDOMIZE” button and then the “START” button for starting the practice session!

- Please click--just once--the “PLAY” button to play the sound pattern! Each sound pattern will be presented 3 times.

- After the sound pattern ends, the cursor (pointer) is going to the textbox automatically. Please type what you heard! Please avoid to guess a correct answer and please do not think too much! The sound(s) may be or may not be a meaningful word. There are three options that you can choose for typing your answer:
 - ✓ If you heard any English word(s) and you are able to identify it, please type it!
 - ✓ If you heard any English sound(s) but you are not sure about its spelling (it may be a nonsense word), please use English alphabet to approximate the sound(s)! I may ask you how to pronounce it later.
 - ✓ If you heard no English word(s)/sound(s), please type “NONE” as your answer!

- Please click once the “NEXT” button for the next sound!

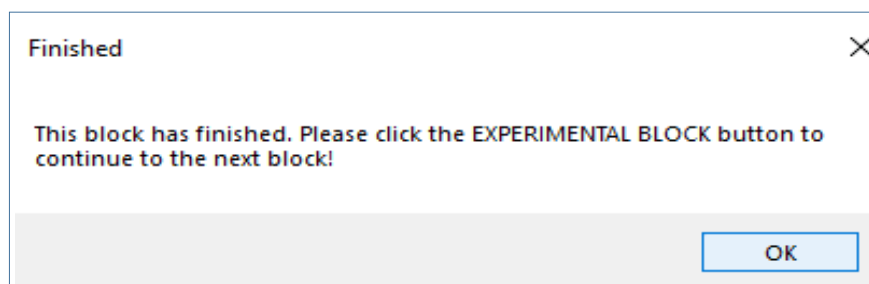
9. Please do the same instructions (5, 6, 7, and 8) until the trial block is finish.
10. After practice session is finished, please click “OK” button for the next session when message box appears!



11. Please select the “MAIN” session, then click the “RANDOMIZE” button and then the “EXPERIMENT BLOCK” button.



12. Please do the same instructions (5, 6, 7, and 8) until the trial block is finished.
13. Please click the “OK” button to the next block of main session experiment when message box appears!



14. Please do the same instructions (10, 11, 12, and 13) until the batch 1 of experiment is finish.

Please do the same instructions (1-14) for the batch 2 of experiment!.

Appendix H. Statistical analysis of Main experiment data

1. Native-English group

- Normality check of English mosaic speech words intelligibility

	Tests of Normality					
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
CompressOMBD20_H	0.222	19	0.014	0.874	19	0.017
CompressOMBD20_W	0.253	19	0.002	0.874	19	0.017
CompressOMBD40_H	0.191	19	0.068	0.917	19	0.101
CompressOMBD40_W	0.213	19	0.023	0.867	19	0.013
PreserveOMBD20_H	0.224	19	0.013	0.850	19	0.007
PreserveOMBD20_W	0.242	19	0.005	0.837	19	0.004
PreserveOMBD40_H	0.262	19	0.001	0.862	19	0.010
PreserveOMBD40_W	0.311	19	0.000	0.776	19	0.001
Stretch2OMBD20_H	0.290	19	0.000	0.803	19	0.001
Stretch2OMBD20_W	0.319	19	0.000	0.783	19	0.001
Stretch2OMBD40_H	0.340	19	0.000	0.811	19	0.002
Stretch2OMBD40_W	0.226	19	0.012	0.866	19	0.012
Stretch4OMBD20_H	0.202	19	0.040	0.911	19	0.076
Stretch4OMBD20_W	0.221	19	0.016	0.911	19	0.076
Stretch4OMBD40_H	0.219	19	0.017	0.885	19	0.026
Stretch4OMBD40_W	0.188	19	0.076	0.886	19	0.027
Stretch8OMBD20_H	0.297	19	0.000	0.617	19	0.000
Stretch8OMBD20_W	0.407	19	0.000	0.647	19	0.000
Stretch8OMBD40_H	0.505	19	0.000	0.445	19	0.000
Stretch8OMBD40_W	0.495	19	0.000	0.460	19	0.000

a. Lilliefors Significance Correction

- Friedman Test between all mosaic speech stimulus types

Test Statistics ^a	
N	19
Chi-Square	188.004
df	19
Asymp. Sig.	0.000

a. Friedman Test

2. Indonesian group

- Normality check of English mosaic speech words intelligibility

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
CompressOMBD20_H	0.299	19	0.000	0.829	19	0.003
CompressOMBD20_W	0.313	19	0.000	0.830	19	0.003
CompressOMBD40_H	0.239	19	0.006	0.886	19	0.027
CompressOMBD40_W	0.197	19	0.052	0.872	19	0.016
PreserveOMBD20_H	0.331	19	0.000	0.776	19	0.001
PreserveOMBD20_W	0.213	19	0.023	0.899	19	0.047
PreserveOMBD40_H	0.190	19	0.069	0.923	19	0.129
PreserveOMBD40_W	0.255	19	0.002	0.873	19	0.016
Stretch2OMBD20_H	0.188	19	0.076	0.923	19	0.126
Stretch2OMBD20_W	0.193	19	0.061	0.917	19	0.098
Stretch2OMBD40_H	0.221	19	0.015	0.852	19	0.007
Stretch2OMBD40_W	0.300	19	0.000	0.796	19	0.001
Stretch4OMBD20_H	0.237	19	0.006	0.859	19	0.009
Stretch4OMBD20_W	0.240	19	0.005	0.818	19	0.002
Stretch4OMBD40_H	0.423	19	0.000	0.549	19	0.000
Stretch4OMBD40_W	0.336	19	0.000	0.704	19	0.000
Stretch8OMBD20_H	0.470	19	0.000	0.536	19	0.000
Stretch8OMBD20_W	0.470	19	0.000	0.536	19	0.000
Stretch8OMBD40_H	0.525	19	0.000	0.362	19	0.000
Stretch8OMBD40_W	0.495	19	0.000	0.460	19	0.000

a. Lilliefors Significance Correction

- Friedman Test between all mosaic speech stimulus types

Test Statistics^a	
N	19
Chi-Square	153.618
df	19
Asymp. Sig.	0.000

a. Friedman Test

3. Chinese group

- Normality check of English mosaic speech words intelligibility

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
CompressOMBD20_H	0.257	20	0.001	0.811	20	0.001
CompressOMBD20_W	0.184	20	0.074	0.881	20	0.018
CompressOMBD40_H	0.200	20	0.035	0.857	20	0.007
CompressOMBD40_W	0.172	20	0.121	0.886	20	0.022
PreserveOMBD20_H	0.296	20	0.000	0.777	20	0.000
PreserveOMBD20_W	0.216	20	0.015	0.916	20	0.083
PreserveOMBD40_H	0.222	20	0.011	0.884	20	0.021
PreserveOMBD40_W	0.210	20	0.021	0.871	20	0.012
Stretch2OMBD20_H	0.288	20	0.000	0.848	20	0.005
Stretch2OMBD20_W	0.283	20	0.000	0.851	20	0.006
Stretch2OMBD40_H	0.250	20	0.002	0.878	20	0.016
Stretch2OMBD40_W	0.202	20	0.032	0.882	20	0.019
Stretch4OMBD20_H	0.238	20	0.004	0.880	20	0.018
Stretch4OMBD20_W	0.250	20	0.002	0.818	20	0.002
Stretch4OMBD40_H	0.340	20	0.000	0.705	20	0.000
Stretch4OMBD40_W	0.238	20	0.004	0.836	20	0.003
Stretch8OMBD20_H	0.394	20	0.000	0.669	20	0.000
Stretch8OMBD20_W	0.413	20	0.000	0.608	20	0.000
Stretch8OMBD40_H	0.538	20	0.000	0.236	20	0.000
Stretch8OMBD40_W	0.450	20	0.000	0.583	20	0.000

a. Lilliefors Significance Correction

- Friedman Test between all mosaic speech stimulus types

Test Statistics^a	
N	20
Chi-Square	129.139
df	19
Asymp. Sig.	0.000

a. Friedman Test