applied sciences

MDPI

Article

Intelligibility of English Mosaic Speech: Comparison between Native and Non-Native Speakers of English

Santi [1,*], Yoshitaka Nakajima [2], Kazuo Ueda [3] and Gerard B. Remijn [4]

[1] Human Science International Course, Kyushu University, Fukuoka 815-8540, Japan
[2] Sound Corporation, Fukuoka 813-0001, Japan; yoshitaka.nakajima@100years.life
[3] Department of Human Science, Faculty of Design/Research Center for Applied Perceptual Science/Research and Development Center for Five-Sense Devices, Kyushu University, Fukuoka 815-8540, Japan; ueda@design.kyushu-u.ac.jp
[4] Department of Human Science, Kyushu University/Research Center for Applied Perceptual Science, Kyushu University, Fukuoka 815-8540, Japan; remijn@design.kyushu-u.ac.jp
* Correspondence: santi.santi.504@s.kyushu-u.ac.jp

Received: 20 August 2020; Accepted: 27 September 2020; Published: 2 October 2020

Abstract: Mosaic speech is degraded speech that is segmented into time × frequency blocks. Earlier research with Japanese mosaic speech has shown that its intelligibility is almost perfect for mosaic block durations (MBD) up to 40 ms. The purpose of the present study was to investigate the intelligibility of English mosaic speech, and whether its intelligibility would vary if it was compressed in time, preserved, or stretched in time. Furthermore, we investigated whether intelligibility differed between native and non-native speakers of English. English (n = 19), Indonesian (n = 19), and Chinese (n = 20) listeners participated in an experiment, in which the mosaic speech stimuli were presented, and they had to type what they had heard. The results showed that compressing or stretching the English mosaic speech resulted in similar trends in intelligibility among the three language groups, with some exceptions. Generally, the intelligibility for MBDs of 20 and 40 ms after preserving/stretching was higher, and decreased beyond MBDs of 80 ms after stretching. Compression also lowered intelligibility. This suggests 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.

Keywords: mosaic speech; temporal resolution; speech intelligibility; compressed speech; stretched speech

1. Introduction

In daily life, we often need to interpret speech that is interrupted or accompanied by other sounds. Various studies have been performed to investigate how humans are able to interpret speech when it is spoken in a noisy environment [1,2], or under reverberation [2,3]. A wealth of research has particularly been performed on temporal aspects of speech processing, by using speech in which parts of the signal were segmented or omitted. An early study on the perception of distorted speech employed speech in which 50-ms portions were alternately played and silenced [4]. 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 [5,6]. Besides periodically interrupted speech, processing of distorted speech has further been investigated with speech that was temporally smeared [7,8] or temporally reversed [9,10]. 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 [11,12]. 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 [13,14]. A study with "pixelated speech" also showed that when speech is divided into segments of 50 ms or shorter, almost the same intelligibility can be obtained as for the original speech [15].

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 [16], articulatory features [17], or syllables [18]. 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 [19]. Neural oscillations with this modulation frequency are thought to be engaged in phonemic processing [20]. Interestingly, the modulation frequency of these neural oscillations corresponds to a temporal window of around 20–33 ms [21], 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 shorter.

Recently, a new type of speech stimulus, called "mosaic speech" [14] has been developed to further study speech processing in general and its temporal aspects in particular. One of the purposes of using "mosaic speech" was to provide an alternative to locally time-reversed speech, because 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 mosaicized Japanese speech, in which the frequency resolution was as fine as a critical bandwidth, Nakajima et al. [14] found that intelligibility was near-perfect (>95%) at shorter block durations of up to 40 ms, similar to the results of studies with locally time-reversed speech. Intelligibility decreased dramatically at longer block durations, from 80 to 320 ms [14]. 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 suitable to simulate the auditory periphery.

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 2.3.2). The initial procedure resembled the procedure to generate noise-vocoded speech [22–25]. First, the original speech signal was separated into several frequency band-pass filters, following the concept of *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, and the total amount of its sound energy was calculated by squaring and adding up instantaneous amplitudes. In the same frequency band-pass filters, a white noise of the same duration as that of the speech signal was generated and the waveforms in these filters were cut into the same temporal segments. Cosine-shaped rise and fall times were used for each of these temporal segments, and the total amount of its sound energy was calculated in the same way as for the speech signal. There was a time-frequency correspondence between the speech signal and the white noise. Each temporally-segmented band noise was amplitude-adjusted to make its total sound energy equal to that of its counterpart in the processed speech signal. By putting all these segmented band noises together on the time-frequency plane, mosaic speech was obtained.

As in the previous study with Japanese mosaic speech [14], OMBD was varied, but only in two steps this time (20 and 40 ms). Although it had been shown that mosaic speech with an OMBD of 40 ms or shorter was almost perfectly intelligible (>95%), so far it was 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 we compress original mosaic blocks of 80 ms, which are not sufficiently understandable, into mosaic blocks of 20 ms, does this improve intelligibility? As an opposing example, if we stretch out original mosaic blocks of 20 ms, which are reasonably intelligible, into mosaic blocks of 80 ms, does intelligibility deteriorate? Thus, our primary goal was to measure the intelligibility of mosaic speech in which OMBD was either compressed, preserved, or stretched in time.

Our secondary goal was to investigate whether the intelligibility of mosaic speech varies with the language background of the listener. Instead of Japanese, as used previously [14], English mosaic speech was used here. Since 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 [26]. 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/ and /f/, and /d/ and /ð/. 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 [27]. 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 [28]. 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 [29]. Whereas near-perfect (>95%) intelligibility of mosaic speech with segments of 40 ms or shorter has been reported for the Japanese language [14], we here investigated whether English mosaic speech, with all its complexity in speech sounds, would also be near perfect at a similar segment duration or not. Apart from a preliminary study in our group [30], no systematic data on the perception of English mosaic speech have been gathered. In the present study, we employed listeners of three different language backgrounds, i.e., native-English listeners, Chinese listeners, and Indonesian listeners, to measure the intelligibility of English mosaic speech.

2. Materials and Methods

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 Chinese and the Indonesian participants were university students who had completed tests of English as a second language. Out of 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). From 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). 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 study 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).)

2.2. Equipment

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.

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 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 and Kjær, 4153, Nærum, Denmark) and a sound level meter (ACO, Type 6240).

2.3. Stimuli

2.3.1. English Word Specifications

Eighty English words in a Consonant-Vowel-Consonant (CVC) structure were used. The words were derived from an English textbook [31] within the category of "content words", which have lexical meanings [32]. 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)/]; [27]. Furthermore, words with two or more possible pronunciations depending on dialect were excluded (e.g., dog; [/dɒg/] or [/dɔg/], [33]. Finally, words that appeared in a homophone or heteronym list were excluded as well [31]. 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/, /ʌ/, /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 [34], 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 (=80% [34]). Words with a vowel /ʊ/ and last consonant /l/ (e.g., *full* and *pull*) were omitted, because they were unintelligible (=10% [34]). 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 [35]. 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 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 1. The 80 CVC-words used in the English mosaic speech experiment.

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	/ɡʌn/		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	/ɡeɪm/
	food	/fu:d/		couch	/kaʊtʃ/		move	/mu:v/		young	/jʌŋ/
C	size	/saɪz/	H	pig	/pɪɡ/	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	/ɡɪv/		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	/ɡʌm/		face	/feɪs/
	check	/tʃek/		judge	/dʒʌdʒ/		shine	/ʃaɪn/		map	/mæp/
	page	/peɪdʒ/		five	/faɪv/		choose	/tʃu:z/		life	/laɪf/

2.3.2. 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 "wav" format with a sampling rate of 44.1 kHz, 16-bit quantization on a mono channel. Each word was recorded three times, from which one was chosen as a stimulus to be used in the experiment. This 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 [36]. Furthermore, we added an empty duration of 10 ms before each word and 5 ms after the end of each word.

In order to generate mosaic speech, we were not able to proceed from any existing mosaicization program for visual images [37]. Since, for sound, an uncertainty principle works between time and frequency, the inverse of time, it is essentially, not just technically, impossible to cut both the horizontal and the vertical axis of the sound spectrogram into pieces very accurately. We therefore carefully constructed an algorithm by which a temporal resolution of 20 ms was secured, and frequency resolution of the narrowest critical band was secured as well, considering the purpose of the present study (see Nakajima et al., 2018 [14]). 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, we separated first the original speech signal 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 [38] (Figure 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 1c. As mentioned earlier, in our previous study [14], it was found that the intelligibility of Japanese mosaic speech was near-perfect (>95%) if the mosaic block duration was 20 or 40 ms. Unpublished data from our preliminary experiment [34] indicated that the intelligibility of mosaicized English speech was on a high plateau of about 70% when the mosaic block duration was 20 or 40 ms. Therefore, we selected OMBDs of 20 and 40 ms for the present study.

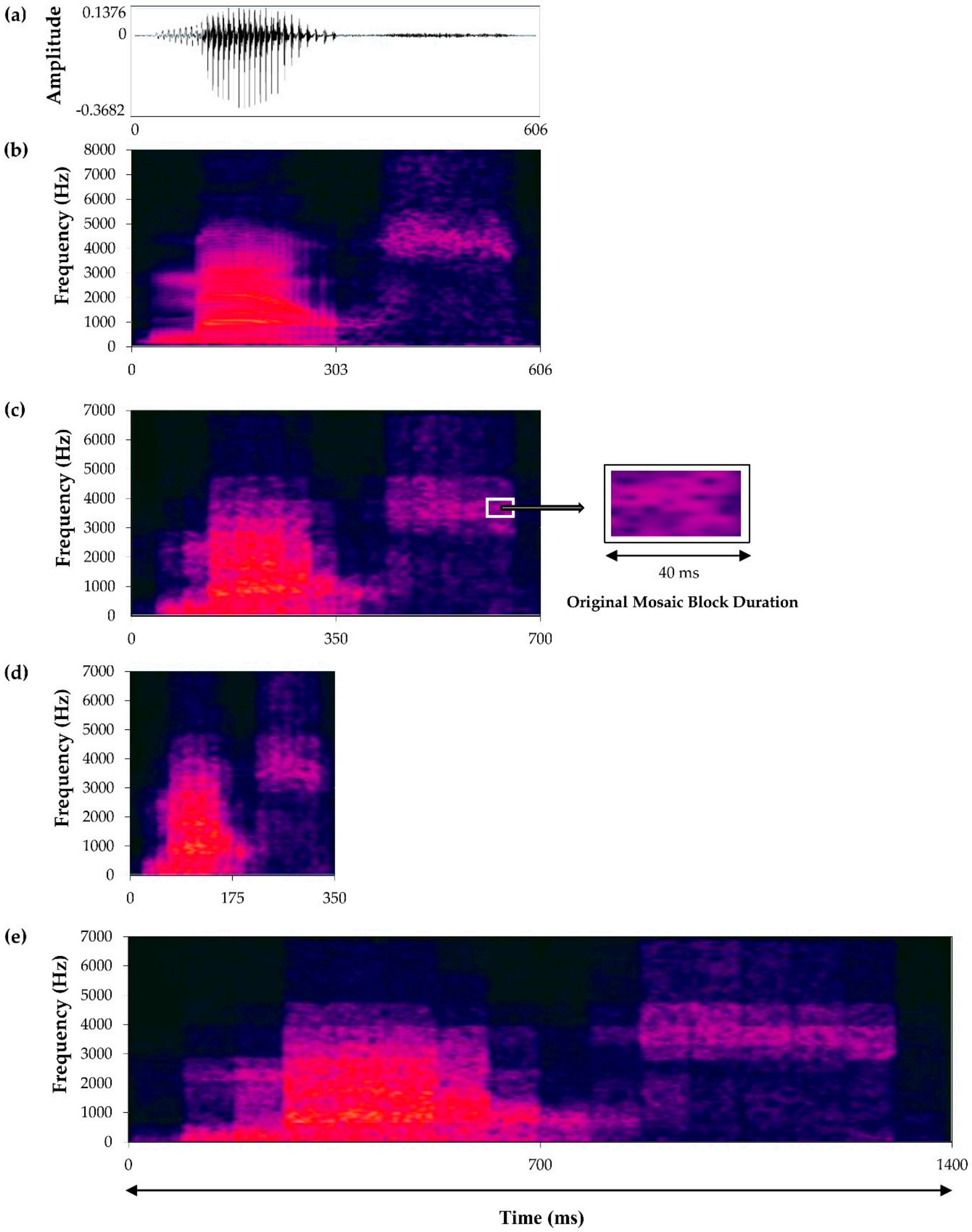

Figure 1. Examples of the mosaic speech stimuli used in this study. (**a**) The waveform and (**b**) the spectrogram of the original speech for the word "mouse", pronounced by a male native-English speaker. (**c**) An example of mosaic speech with an original mosaic block duration (OMBD) of 40 ms. Each 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.

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 processed speech signal. Mosaic speech was obtained by putting all these amplitude-adjusted segmented band noises together on the time-frequency plane.

For the present experiment, we compressed (Figure 1d), preserved, or stretched (Figure 1e) the OMBD by reducing, keeping, or increasing the number of samples for each mosaic block. The OMBD was compressed into half (0.5 × OMBD), preserved (1 × OMBD), or stretched by a factor of 2, 4, or 8 (2 × OMBD, 4 × OMBD, 8 × OMBD). The resulting duration was called "Mosaic Block Duration" (MBD); the shortest MBD was 10 ms (0.5 × OMBD of 20 ms), and the longest was 320 ms (8 × 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. Mosaic speech block durations used in the experiment.

	OMBD: 20 ms		**OMBD: 40 ms**	
MBD after Compressing/Preserving/Stretching	**Mosaicizing Phase Type**			
	Half (10 ms)	**Whole (20 ms)**	**Half (20 ms)**	**Whole (40 ms)**
Compressed (OMBD × 0.5)	10	10	20	20
Preserved (OMBD × 1)	20	20	40	40
Stretched 2 (OMBD × 2)	40	40	80	80
Stretched 4 (OMBD × 4)	80	80	160	160
Stretched 8 (OMBD ×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, we added another portion of empty duration to make it a half or a whole length of the OMBD, as indicated in Table 2 ("Half" and "Whole"). By using two different lengths of the total added duration (10 ms or 20 ms for the OMBD of 20 ms; 20 ms or 40 ms for the OMBD of 40 ms), we could 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.

Besides the words transformed into mosaic speech, the same original words were also presented as control stimuli to check whether the participants, especially the Indonesian and the Chinese participants, knew these words.

2.4. Procedure

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 1). Each group was assigned to a different mosaic speech stimulus type, and this assignment was different among participants. 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 min 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.

2.5. Data Processing

The Friedman two-way analysis of variance by ranks [39] was performed to analyze the main effect of compressing or stretching the OMBD on the intelligibility scores for each language group. Although this was not the main purpose of the study, the Wilcoxon signed-rank test [39] 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 type was found, for convenience the scores were collapsed. The Wilcoxon signed-rank test [39] was used to further analyze the effect of compressing, preserving, or stretching the 20-ms or 40-ms OMBD within each language group by making multiple comparisons. Post-hoc Holm-Bonferroni correction [40,41] 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 language group. 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).

3. Results

3.1. Intelligibility Comparisons between Original Speech and Mosaic Speech

Intelligibility (i.e., word identification accuracy) was obtained by counting the number of correct answers of all participants for all stimuli. Figure 2 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.

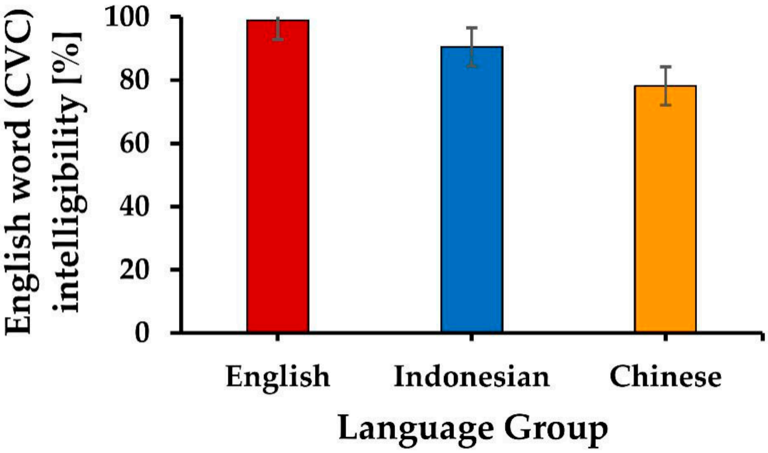

Figure 2. Word identification accuracy (intelligibility) for original speech for each language group (English, *n* = 19; Indonesian, *n* = 19; Chinese, *n* = 20). Error bars indicate standard error of means.

The intelligibility scores for the mosaic speech stimuli are shown in Figure 3. Since the results of the Wilcoxon signed-rank test [39] 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 the present study, for convenience the scores were collapsed for subsequent analyses. Figure 3 shows that mosaicizing 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. An obvious reason for this is that when original speech is mosaicized, its signal degrades both in the temporal and the frequency dimension.

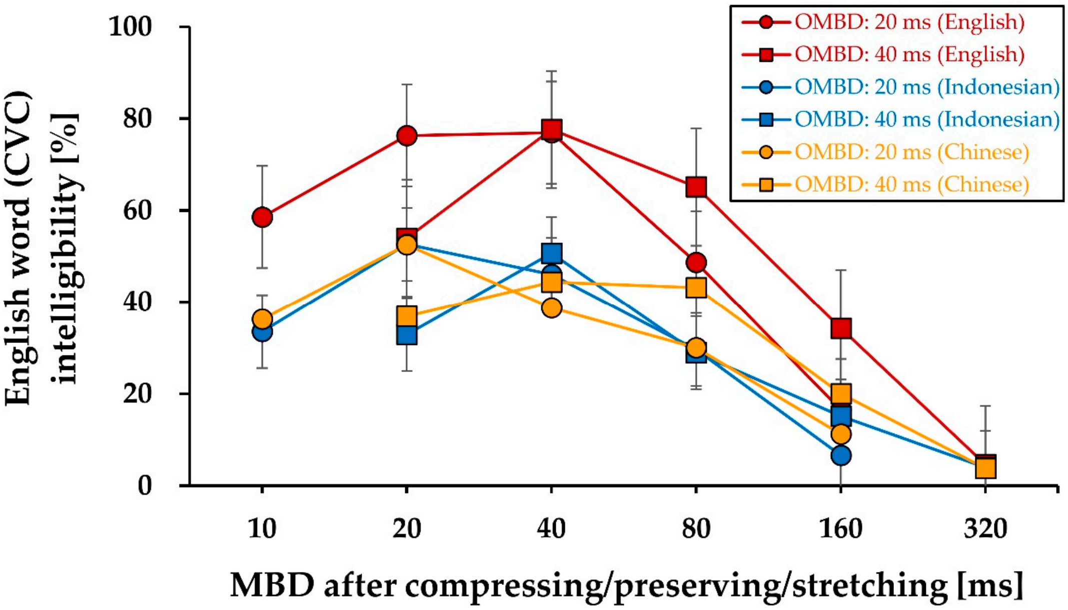

Figure 3. English word identification accuracy (intelligibility) for mosaic speech as functions of MBD after compressing/preserving/stretching (English, $n = 19$; Indonesian, $n = 19$; Chinese, $n = 20$). The data for half-phase and whole-phase types are collapsed. Error bars indicate standard errors.

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

3.2.1. 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 the MBD of 40 ms after preserving or stretching, and decreased as the MBD was shorter or longer. When the OMBDs were compressed, the intelligibility decreased by about 20%. 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 [39] 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 [39] were used to follow up this finding, and a Holm-Bonferroni correction [40,41] was performed after that. In detail, 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 ($p < 0.01$).

3.2.2. Non-Native Listeners of English

The intelligibility of English mosaic speech for the Indonesian and the Chinese listeners was highest 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 listeners. The intelligibility decreased by about 19% for the MBDs after compressing for the Indonesian and the Chinese group. The intelligibility decreased sharply when the OMBDs were stretched into MBDs of 80 ms or longer, except in the Chinese group when the OMBD of 40 ms was stretched by a factor of 2.

By the same statistical method as described above, here too we found significant differences in intelligibility between mosaic speech types, for the Indonesian participants (n = 19, k = 19, χ^2 = 153.618, p < 0.001) and the Chinese participants (n = 20, k = 19, χ^2 = 129.139, p < 0.001). In detail, for the Indonesian group, the intelligibility of mosaic speech with the MBD of 20 ms after preserving was significantly higher than that of the same 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 same MBDs after stretching (80, 160, or 320 ms; p < 0.01). Meanwhile for the Chinese group, for both OMBDs, there were no significant differences in intelligibility between the preserved and the compressed or the stretched × 2 stimuli (p > 0.05), nor between the compressed and the stretched × 2 or × 4 stimuli (p > 0.05). Both the Indonesian and 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.

Although the Indonesian and the Chinese listeners overall had significantly lower intelligibility scores than the native-English listeners, as can be seen in Figure 3, there were similar trends in intelligibility across the three language groups.

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-ms and a 40-ms OMBD contained blocks of the same duration, and we also wanted to know whether the word intelligibility for stimuli with the same MBD would be similar or not. 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 MBD of 20 ms (p = 0.004) after preserving for the native-English listeners. Between the stimuli of the same MBD of 40, 80, or 160 ms, no significant differences in intelligibility were found (p > 0.05).

For the Indonesian group, there was no significant effect of compression for the stimuli with an MBD 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).

4. Discussion and Conclusions

In order to investigate temporal aspects of speech processing, we performed an intelligibility test of mosaicized 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. Listeners with three different language backgrounds (native-English, Indonesian, and Chinese) were employed to determine the intelligibility of mosaicized 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 or 40-ms MBDs (intelligibility score 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 stretched by a factor of 4 or 8, mosaic speech was basically unintelligible. 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 preserved or stretched.

A second goal of this study was to investigate whether the effects of compressing, preserving or stretching 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 ms 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 ms 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 our preliminary experiment [34].

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. This agrees with Nakajima et al.'s (2018) [14] 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 [42]). They suggested that the temporal resolution of about 40 ms is necessary to perceive motion in general. Furthermore, as mentioned in the Introduction, 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 [19,21]. 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.

Before considering practical implications of mosaic speech, the following needs to be addressed first. In the previous study with Japanese mosaic speech [14], the intelligibility was near-perfect (>95%) for native-Japanese listeners. However, in the present study with English mosaic speech, the native-English group reached intelligibility scores of only around 75% on average. It is possible that the complexity of the English speech sounds, as discussed in the Introduction, also affected intelligibility. Therefore, we aim 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 [43]. 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 Japanese mosaic speech, we here used words and not sentences. Identifying a word in isolation is not as easy as identifying a word in a sentence context [44]. It is thus necessary to investigate what is specific for the 40-ms window.

Author Contributions: Conceptualization, S., and Y.N.; methodology, S. and Y.N.; software, S. and Y.N.; validation, S., Y.N. and G.B.R.; formal analysis, S., Y.N.; investigation, S., Y.N.; resources, S. and Y.N.; data curation, S., Y.N. and G.B.R.; writing—original draft preparation, S. and G.B.R.; writing—review and editing, S., Y.N., K.U. and G.B.R.; supervision, Y.N. and G.B.R.; project administration, Y.N., K.U. and G.B.R. All authors have read and agreed to the published version of the manuscript.

Funding: This study and the APC were funded by JSPS KAKENHI Grant Numbers JP17H06197 and JP19H00630.

Acknowledgments: We wish to thank 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 and João Paulo Cabral for their support in data collection, and Alexandra Wolf for her comments on an initial version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yoo, S.; Boston, J.; El-Jaroudi, A.; Li, C. Speech signal modification to increase intelligibility in noisy environments. *J. Acoust. Soc. Am.* **2007**, *122*, 1138–1149. [CrossRef] [PubMed]
2. Crespo, J.; Henriks, R. 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), Florence, Italy, 4–9 May 2014; pp. 910–914.
3. Dong, H.; Lee, C. Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering. *EURASIP J. Audio Speech Music Process.* **2018**, *3*, 1–13. [CrossRef]
4. Miller, G.A.; Licklider, J.C.R. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.* **1950**, *22*, 167–173.
5. Fairbanks, G.; Kodman, F., Jr. Word intelligibility as a function of time compression. *J. Acoust. Soc. Am.* **1957**, *29*, 636–641. [CrossRef]
6. Shafiro, V.; Sheft, S.; Risley, R. The intelligibility of interrupted and temporally altered speech: Effects of context, age, and hearing loss. *J. Acoust. Soc. Am.* **2016**, *139*, 455–465. [CrossRef]
7. Drullman, R.; Festen, J.M.; Plomp, R. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* **1994**, *95*, 2670–2680. [CrossRef]
8. Drullman, R.; Festen, J.M.; Plomp, R. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* **1994**, *95*, 1053–1064. [CrossRef]
9. Kellogg, E.W. Reversed speech. *J. Acoust. Soc. Am.* **1939**, *10*, 324–326.
10. Meyer-Eppler, W. Reversed speech and repetition systems as means of phonetic research. *J. Acoust. Soc. Am.* **1950**, *22*, 804–806.
11. Steffen, A.; Werani, A. Ein Experiment zur zeitverarbeitung bei der Sprachwahrnehmung. In *Sprechwissenschaft & Psycholinguistik*; Kegel, G., Arnhold, T., Dahlmeier, K., Schmid, G., Tischer, B., Eds.; Westdeutscher Verlag: Opladen, Germany, 1994; Volume 6, pp. 189–205.
12. Saberi, K.; Perrott, D.R. Cognitive restoration of reversed speech. *Nature* **1999**, *398*, 760.
13. Ueda, K.; Nakajima, Y.; Ellermeier, W.; Kattner, F. Intelligibility of locally time-reversed speech: A multilingual comparison. *Sci. Rep.* **2017**, *7*, 1782. [CrossRef] [PubMed]
14. Nakajima, Y.; Matsuda, M.; Ueda, K.; Remijn, G.B. Temporal resolution needed for auditory communication: Measurement with mosaic speech. *Front. Hum. Neurosci.* **2018**, *12*, 149. [CrossRef] [PubMed]
15. Schlittenlacher, J.; Staab, K.; Çelebi, Ö.; Samel, A.; Ellermeier, W. Determinants of the irrelevant speech effect: Change in spectrum and envelope. *J. Acoust. Soc. Am.* **2019**, *145*, 3625–3632. [CrossRef] [PubMed]
16. Liberman, A.M.; Mattingly, I.G. The motor theory of speech perception revised. *Cognition.* **1985**, *21*, 1–36. [CrossRef]
17. Stevens, K.N. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* **2002**, *111*, 1872–1891. [CrossRef]
18. Greenberg, S.; Arai, T. What are the essential cues for understanding spoken language? *IEICE Trans. Inf. Syst.* **2004**, *E87-D*, 1059–1070. [CrossRef]
19. Giraud, A.L.; Poeppel, D. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat. Neurosci.* **2012**, *15*, 511–517. [CrossRef]
20. Ding, N.; Patel, A.D.; Butler, H.; Luo, C.; Poeppel, D. Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **2017**, *81*, 181–187. [CrossRef]

21. Chait, M.; Greenberg, S.; Arai, T.; Simon, J.Z.; Poeppel, D. Multi-time resolution analysis of speech: Evidence from psychophysics. *Front. Neurosci.* **2015**, *9*, 214. [CrossRef]
22. Shannon, R.V.; Zeng, F.-G.; Kamath, V.; Wygonski, J.; Ekelid, M. Speech recognition with primarily temporal cues. *Science* **1995**, *270*, 303–304. [CrossRef]
23. Smith, Z.M.; Delgutte, B.; Oxenham, A.J. Chimaeric sounds reveal dichotomies in auditory perception. *Nature* **2002**, *416*, 87–90. [CrossRef] [PubMed]
24. 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. *J. Acoust. Soc. Am.* **2015**, *138*, 1561–1569. [CrossRef] [PubMed]
25. Kishida, T.; Nakajima, Y.; Ueda, K.; Remijn, G. Three factors are critical in order to synthesize intelligible noise-vocoded Japanese speech. *Front. Psychol.* **2016**, *7*, 517. [CrossRef] [PubMed]
26. Wells, J.C. *Accent of English*; Cambridge University Press: Cambridge, UK, 1982.
27. Carley, P.; Mees, I.M.; Collins, B. Basic Concepts. In *English Phonetics and Pronunciation Practice*; Routledge: Abingdon, London, UK; New York, NY, USA, 2018; pp. 1–2.
28. Volín, J.; Skarnitzl, R. Foreign Accents and English in International Contexts. In *the Pronunciation of English by Speakers of Other Languages*; Cambridge Scholars Publishing: Newcastle Upon Thyne, UK, 2018; pp. 1–2.
29. Wenanda, D.; Suryani, S. Analisis Kesalahan Berbahasa Inggris pada Tataran Fonologis. In *Prosodi: Jurnal Ilmu Bahasa dan Sastra*; Department of English, University of Trunojoyo: Madura, Indonesia, 2016; Volume X, Nomor 2; pp. 145–155.
30. Kojima, K.; Nakajima, Y.; Ueda, K.; Remijn, G.B.; Elliott, M.A.; Arndt, S. 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, Fukuoka, Japan, 22–26 October 2017; p. 127.
31. Kress, J.E.; Fry, E.B. *The Reading Teacher's: Book of List*, 6th ed.; Jossey-Bass: San Francisco, CA, USA, 2016; pp. 21–171.
32. Richards, J.C.; Schmidt, R.W. *Longman Dictionary of Language Teaching & Applied Linguistics*, 4th ed.; Routledge: Abingdon, London, UK; New York, NY, USA, 2010; p. 126.
33. Wells, J.C. *Longman Pronunciation Dictionary*, 3rd ed.; Pearson: London, UK, 2008.
34. Santi, S.; Nakajima, Y.; Ueda, K.; Remijn, G.B. 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, Antalya, Turkey, 30 October–2 November 2019; p. 35.
35. Wells, J.C. *Longman Dictionary of Contemporary English*, 6th ed.; Pearson: London, UK, 2014.
36. Cambridge Dictionary. Available online: https://dictionary.cambridge.org/ (accessed on 26 July 2019).
37. Harmon, L.D. The recognition of faces. *Sci. Am.* **1973**, *229*, 71–82. [CrossRef] [PubMed]
38. Fastl, H.; Zwicker, E. Critical Bands and Excitation. In *Psychoacoustics: Facts and Models*, 3rd ed.; Springer: New York, NY, USA, 2007; pp. 149–172.
39. Field, A. Non-parametric Tests. In *Discovering Statistics Using SPSS*, 3rd ed.; Sage Publication: London, UK, 2009.
40. Abdi, H. Holm's Sequential Bonferroni Procedure. In *Encyclopedia of Research Design*; Salkind, N.J., Ed.; Sage: Thousand Oaks, CA, USA, 2010; pp. 573–577.
41. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
42. Read, P.; Meyer, M.P. Cinematographic Technology. In *Restoration of Motion Picture Film*; Elsevier: Amsterdam, The Netherlands; Butterworth-Heinemann: Oxford, UK; Boston, MA, USA, 2000; pp. 9–45.
43. Fontan, L.; Ferrané, I.; Farinas, J.; Pinquier, L.; Tardieu, J.; Magnen, C.; Gaillard, X.A.; Füllgrabe, C. Speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *J. Speech Lang. Hear. Res.* **2017**, *60*, 2394–2405. [CrossRef]
44. Marslen-Wilson, W.D. Function and process in spoken word-recognition. In *Attention and Performance X: Control of Language Processes*; Bouma, H., Bouwhuis, G., Eds.; Erlbaum: Hillsdale, NJ, USA, 1984; pp. 125–150.