Chen *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:19
DOI 10.1186/s13636-016-0097-6

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

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

Bohan Chen[1,3,4*], Norihide Kitaoka[2,4] and Kazuya Takeda[1,3,4]

Abstract

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

Keywords: Subtle prosodic cues, Prosody information transmission efficiency, Voice morphing, Eye tracking, Objective similarity measure

1 Introduction

Language comprehension involves a complex interaction between the transmitted message and the receiver's background knowledge and experiences [1]. As a result of this complexity, differences in representation styles can clearly influence the efficiency of our language comprehension process. For example, the inversion of subject and object in passive sentences makes these sentences more difficult for listeners to understand than sentences with the same meaning expressed using active voice, for both positive and negative sentences [2]. Listeners also have difficulty interpreting "garden path" sentences, i.e., grammatically correct sentences which have meanings different from those that a listener would normally expect. For example, "The dog that I had really loved bones," and "I told her children are noisy." Such sentences are considered to be evidence of our sequential reading process (i.e., one word read at a time) [3].

Schema theory suggests that presenting messages in style that is familiar to the recipient improves comprehension efficiency, because when a receiver has relevant background knowledge, he or she can free up more working memory for analysis and interpretation of the message [4, 5]. Researchers have found evidence to support the theory that both lexical and prosodic familiarity increase the efficiency of our language comprehension. Use of familiar topics has been found to help foreign language learners improve their performance on reading comprehension tasks, no matter which second language they are learning [6] or what their native language is [7]. Moreover, the facilitative effect of comprehension on language-related tasks is revealed in simple nativization drills, such as the changing of character and location names into native ones (e.g., when a Japanese English learner replaces "Barack Obama lives in Washington D.C." with "Shinzo Abe lives in Tokyo") [8]. Studies also show that familiarity with the speaker's speech characteristics, such as the speaker's accent, also have a positive influence on our listening comprehension, for both native and non-native listeners [9, 10].

*Correspondence: bohan.chen@g.sp.m.is.nagoya-u.ac.jp
[1]Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan
[3]Institute of Innovation for Future Society (MIRAI), Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan
Full list of author information is available at the end of the article

In most of the cases mentioned above, familiarity also involves self-similarity (i.e., we are familiar with our own accent, capital, president, etc.). Thus, it seems that self-similarity is a factor related with high-efficiency communication. However, most of these researches employed second language learner as their participants, there is still lack of evidence to show whether subtle prosodic cues significantly influence our listening comprehension. It is important to us because we aim to find a way to predict and achieve (through speech synthesis) high-efficiency speech communication, if subtle prosodic cues cannot significantly influence our comprehension, the idea can hardly be applied. Thus, in previous research we have tried to use speaker self-similarity as a predictor of information transmission quality in dialogues [11]. We investigated the relationship between similarity in spectral envelope features, prosodic features and lexical features of speakers and listeners and the quality of information transmission during map task dialogues. Prosodic and lexical similarity were found to be correlated with information transmission quality, and spectral envelope similarity was also found to have a weak but significant correlation with map task performance. These results surprised us, because it is well known that the perception of one's own voice involves a mixture of air conduction and bone conduction [12], meaning that our perception of our recorded voice differs from our daily perception of our own voices. In fact, we rarely perceive our own voice to be familiar when heard on a recording. Our previous research thus suggests that it is reasonable to assume that we find our own recorded voices more familiar than the recorded voices of others. However, it is still unclear whether the familiarity of subtle prosodic cues, such as fundamental frequency, have a facilitative effect on comprehension efficiency. It is also unclear whether self-similarity influences communication efficiency when subjects hear synthesized voices as it does when communicating face-to-face with real people. Since the correlation is too weak to reach a definitive conclusion, we decided to design an experiment to investigate the effect of voice similarity on comprehension efficiency by observing comprehension when messages are presented at different levels of voice similarity.

Therefore, in this study we designed a behavioral experiment to answer the following questions:

- Does similarity in the speech characteristics of the information sender and information receiver result in higher information transmission efficiency?
- Do subtle acoustic cues, such as spectral envelop, have any influence on the efficiency of information transmission?

This paper is organized as follows. After a description of our experimental method, we describe our experimental procedure, report our experimental results, and discuss their implications. We then end the paper with our conclusions and a discussion of our future research.

2 Method

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

2.1 Material

We employed spoken Japanese phrases with right-branching (RB) vs. left-branching (LB) ambiguities as our experimental material. Figure 1a shows an example[1]. In Japanese sentences such as "akai/hoshi no/nekutai" ("red (adjective phrase)/star (first noun phrase)/necktie (second noun phrase)") can be interpreted, as in English, as either "the red necktie with stars" or "the necktie with red stars." It is RB when the second phrase (the first noun phrase) should first be combined with the third phrase (the second noun phrase) (i.e., "the red necktie with stars"), and LB when the second phrase should first be combined with the first phrase (i.e., "the necktie with red stars"). These two phrases are identical in spelling and phonetic pronunciation but can be distinguished by subtle prosodic cues [13]. No clear downstepping [2] "↘" from the first phrase to the second phrase, followed by downstepping "↘" from the second phrase to the third phrase suggests the right-branching meaning (the red necktie with stars), while clear downstepping "↘" from the first phrase to the second phrase, followed by moving up of pitch "↗" from the second phrase to the third phrase suggests the left-branching meaning (the necktie with red stars)[3]. A longer pause between the first and second phrases also indicates the RB meaning, while a longer pause between the second noun and its particle ("no"), inside the second phrase, indicates the LB meaning. A third prosodic cue is called "final segment duration," which is the duration of the final vowels in the different phrases. When the RB meaning is intended, there is longer final segment duration in the first phrase, while longer final segment duration in the second phrase implies the LB meaning (also see Fig. 2a, b).

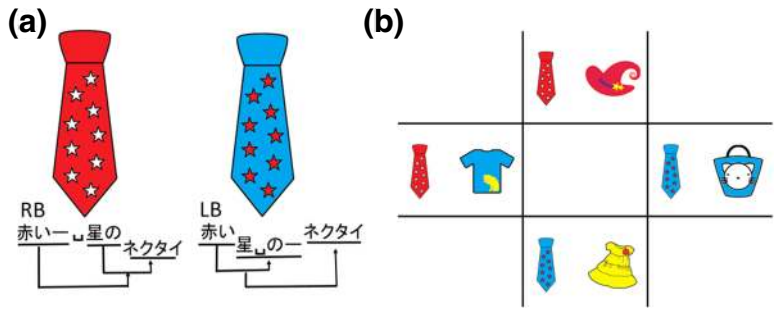


Fig. 1 Example of experimental items. **a** Example of RB vs. LB ambiguity items used for recording; both of the pitcured items can be referred to as "akai hoshi no nekutai" in Japanese ("*red star necktie*" in English). RB prosodic cues: (1) No clear downstepping from the first phrase to the second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

2.2 Voice morphing

Morphing techniques have been developed to change one stimulus object (e.g., an image) into another with a seamless transition. Since morphing techniques can enrich the level of stimulus without salient loss of naturalness, they have been used in many facial image-related experiments, such as those involving facial recognition [14] and attractiveness perception [15]. TANDEM-STRAIGHT [16] is a speech analysis, modification and re-synthesis framework, which can similarly deconstruct a speech signal based on the source-filter model. TANDEM-STRAIGHT extracts the F0 and aperiodicity of the input speech signal as the source parameters. The signal's spectrogram information was used together with its F0 to obtain the filter parameters. While morphing, the weighted average of all the parameters from the two source signals, which also included mapping information in the time and frequency domains, were used to re-synthesize the voice, based on the source-filter model[4] (see Fig. 3a). TANDEM-STRAIGHT can generate naturally sounding voices, allowing acoustic researchers to apply morphing techniques in their experiments in order to investigate the perception of paralinguistic and non-linguistic information in voices, such as the perception of gender [17] and speaker identification [18].

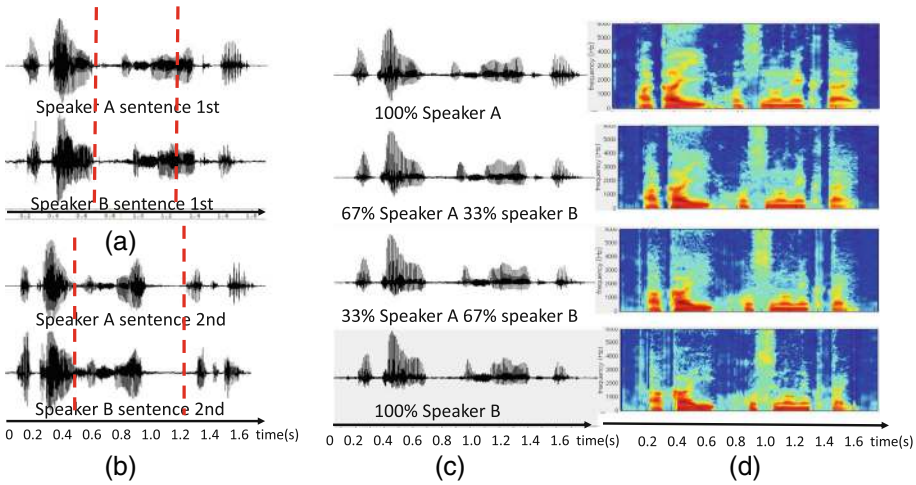
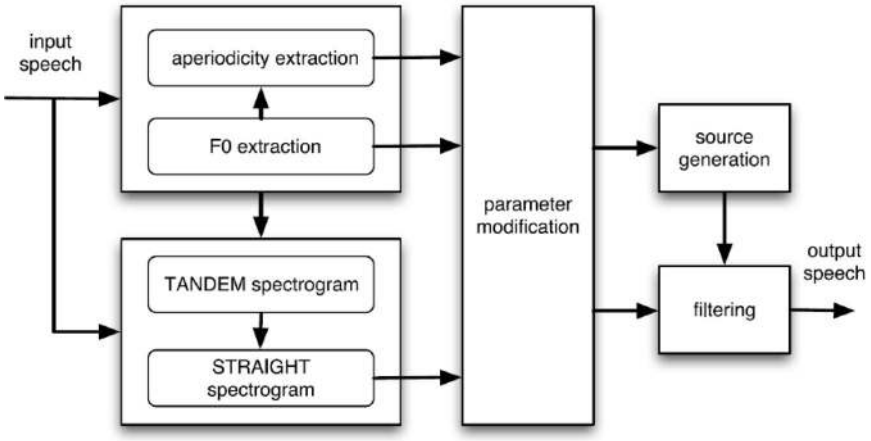
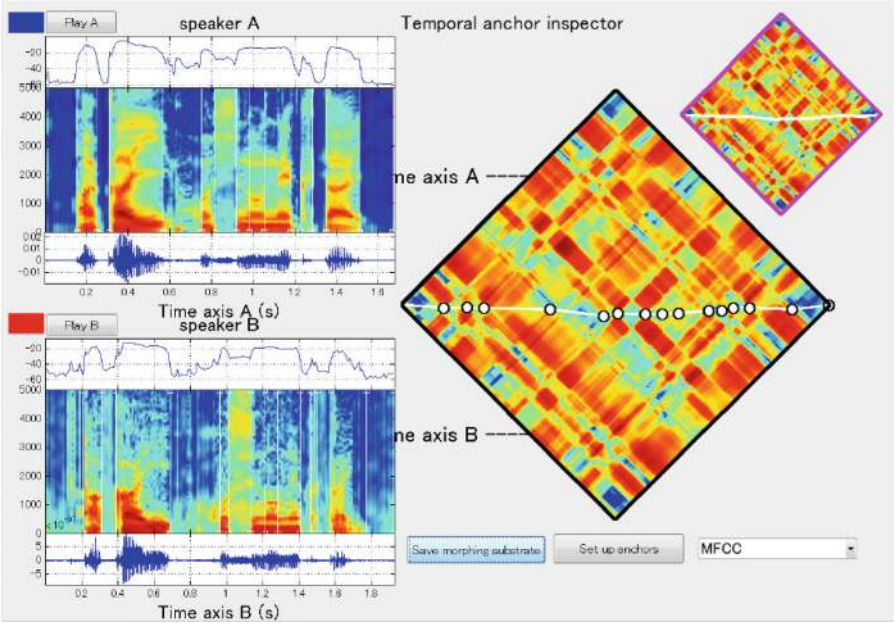


Fig. 2 Examples of different waveforms. **a** Original waveforms of the phrase "the *red necktie with stars*" (RB) as read by different participants. **b** Original waveforms of "the *necktie with red stars*" (LB) as read by different participants. The *dashed lines* show the boundaries of each phrase in the upper sentence. **c** Synthesized waveforms when morphing the waveforms (**a**) together under different morphing conditions. **d** Spectrogram information of waveforms shown in (**c**)



(a)



(b)

Fig. 3 TANDEM-STRAIGHT toolbox for voice morphing. **a** Flow chart of TANDEM-STRAIGHT for voice synthesis. TANDEM-STRAIGHT extracts the F0 and aperiodicity of the input speech signal as the source parameters. The signal's spectrogram information was used together with its F0 to obtain the filter parameters. While morphing, the weighted average of all the parameters from the two source signals (also included other information such as mapping information in time and frequency domains) were used to re-synthesize the voice, based on the source-filter model. **b** Time anchor panel for voice morphing. The diagonally oriented square is the distance matrix of signal A and signal B. The *white circles* in the distance matrix are anchored points, which can be determined manually. *White lines* between anchored points show the aligned frames

After the participants' voices were recorded reading the Japanese RB vs. LB ambiguous phrases, we randomly paired participants with a stranger[5] and used the TANDEM-STRAIGHT toolbox to morph their original voices into four transitional levels of similarity using manually anchored start and end points of each syllable. The starting point and ending point of each syllable were aligned manually (see Fig. 3b, the *white circles* are the anchored points). The morphing conditions were as follows: 100% speaker A's voice, 67% speaker A's voice mixed with 33% speaker B's voice, 33% speaker A's voice mixed with 67% speaker B's voice, and 100% speaker B's voice. As the synthesized voices still sound somewhat artificial, to compensate for this, voices were synthesized using TANDEM-STRAIGHT even for the 100 and 0% similarity conditions. Figure 2c, d show the morphed waveforms and spectrum based on the waveforms shown in Fig. 2a, respectively. And we can see

that they are very similar to each other in timing and intensity.

2.3 Objective similarity measures

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

2.3.1 Spectrum similarity measure

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

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

as its other elements[6]. $D(i,j)$ is the entry of the local distance matrix and $C(i,j)$ is the entry of the cost matrix. Thus, the final entry in the cost matrix (e.g., $C(I,J)$) is the optimum global alignment cost. The optimum mapping path between the two input vectors can also be found by backtracking the optimum path of each node. In this paper, MFCC distance is used to compute the distance between each pair of spectra (one for partner A and one for partner B) for a given phrase (e.g., "red star necktie") so that we can obtain a distance matrix.

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

2.3.2 Pitch contour similarity measure

The weighted correlation proposed in [19] is used for measuring similarity between a pair of pitch contours. After aligning two speech segments using DTW (as explained in the previous subsection), their pitch contour similarity is then computed using the following formula:

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

where $f_A(i)$ and $f_B(i)$ represents the log F_0[7] value of speaker A and B in the ith aligned frame, respectively, m_A and m_B represent the mean log F_0 of speaker A and B in the current speech segment, respectively. I represents the number of frames in the aligned sequence, and $w(i)$ is the weighting factor, based on the frame signal power[8].

2.3.3 Duration similarity measure

The absolute mean difference between anchored intervals (in this case, representing syllable and pause duration) is used for measuring similarity between two sets of anchored speech. After anchoring the start point and end point of each syllable manually, duration similarity is measured using the following formula

$$D_{S_A,S_B} = \frac{1}{N-1} \sum_{s=1}^{N-1} |S_A(s) - S_B(s)|, \quad (3)$$

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

2.4 Procedure

Our experiment was divided into two phases. In the recording phase, participants were shown 13 pairs of pictures. The two pictures in each pair were different, but could be described using the same lexically ambiguous phrase, depending on whether the RB or LB reading was used. They were asked to describe each picture in Japanese twice, using their own natural speaking style, by reading the supplied ambiguous phrase. Example pictures and an example description are shown in Fig. 1a. They were recorded in a sound-proof booth at 48,000 Hz with 20 bits sampling. Participants were then randomly paired with a stranger participant, and TANDEM-STRAIGHT was used to morph their voices with the voices of their partners.

In the second phase of the experiment, a listening comprehension experiment was performed about 1 week later. After completing two unambiguous warm-up trials, the only aim of which was to make sure that the participants

understood what they should do during the experiment, participants listened to the previously recorded ambiguous phrases (in which their voices had been re-synthesized and morphed) while viewing pictures (1024 × 768 pixels) shown on a visual display (see Fig. 1b). Participants were asked to identify which target/image they heard described as quickly as possible by pressing one of four arrow keys on the keyboard. Note that participants listened to exactly the same phrases as their randomly paired stranger partner, the only difference being that the self-similarity conditions differed (i.e., one participant's voice was the "other's person's voice" for their partner, and vice versa).

During the experiment, the eye movements of the participants were tracked with a Tobii X2-30 eye tracker at a sampling frequency of 30 Hz. The targets were pictures of pairs of items, all of which had been seen by the participants during the recording phase of the experiment. Participants were shown a target, which was a set of four pairs of pictures. We called the item on the left of each pair the "first item" (i.e., the necktie in Fig. 1b), and the item on the right of each pair was called the "second item." The first item in each pair was the subject of the ambiguous phrase, while the second item was unique and was described without ambiguity. We included these "second items" because the prosodic differences between the descriptions of pairs of ambiguous options is very subtle. Based on previous research, even when listeners hear their own recorded voices, they can only achieve a comprehension accuracy of about 70%. By adding a unique "second item", we are able to better distinguish between confused responses (when the listener does not know which target is being described) and incorrect responses (when the listener presses the wrong key by mistake). Each set of four pairs of pictures included two pairs with correct first items and two pairs with first items, which could be easily mistaken for the correct items due to RB vs. LB ambiguity.

The listening comprehension experiment involved a total of 60 similar trials (i.e., we randomly selected 15 trials from the 26 in each morphing condition for each pair of participants). Figure 4 shows an example of one trial. Participants were asked to select the correct pair of items based on the phrase they heard by using a keyboard. The phrases were a combination of the participants' morphed voices ("first item" and "second item") in the same morphing condition. As shown in Fig. 4, each trial was divided into four logical stages. The first stage was a 5-s preparation stage, in which the set of four picture pairs was shown without any sound. The second stage ran from the beginning of the description of the first item (the item on the left) to the end of the description of the first item. In the third stage, the participants heard the word "to" (pronounced like the word "toe" in English, which means "and" in Japanese) and then a 0.3-s pause. The fourth and final stage spanned the period from the beginning of the description of the second item until the participant's response via the keyboard[9].

2.5 Participants

Twenty-eight male, native Japanese-speaking college students were recruited as participants[10]. Data collected from four of the participants was removed from analysis either because of experimental error (the participants misunderstood the task) or due to data recording error (50% of their eye movement data was lost). Thus, the study was conducted using data collected from 24 participants[11].

3 Results

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

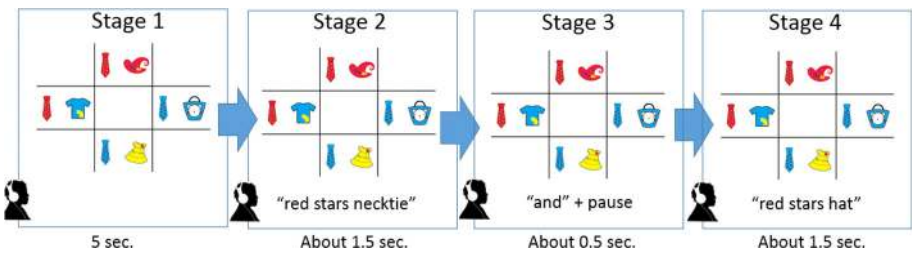


Fig. 4 Experimental procedure. The experimental procedure was divided into four stages. In the first stage, only visual information is presented. During the second stage, information about the ambiguous item is presented. In the third stage, the word "to" (which corresponds to "and" in English) is heard, followed by a 0.3-s pause. In the fourth stage, information about the unique item is presented. The participant's comprehension of the ambiguous information is considered to occur during the second and third stage

test was applied for pairwise comparisons when ANOVA shows that there is a significant difference in experimental performance.

We further divided the "stranger's voice" data into "strangers with voices similar to the listener's own voice" and "strangers with voices dissimilar to the listener's own voice" based on the objective similarity measures, which can be considered to be an extension of the original morphing experiment. We set the 33 and 67% of all the data as thresholds for "similar stranger" and "dissimilar stranger," respectively. Participant pairs whose average objective similarity measure was higher or lower than these thresholds were considered to be a "similar stranger" or "dissimilar stranger," respectively. Further, ANOVA analysis was applied using the "similar stranger" and "dissimilar stranger" categories as an additional "between subjects" factor. Because we were afraid that similarity of pitch and duration of utterances within a participant pair could change (i.e., some utterances could sound similar while other utterances sounded dissimilar), for the purpose of analysis, both pitch and duration similarities were treated as both a "between subjects" factor and a "within subjects" factor (i.e., they were analyzed twice)[12]. Also note that there were only tiny differences in prosodic expression between paired participants. The mean and variance of the mean differences in syllable and pause duration were 44.4 ms and 378.04(ms)^2, respectively. The mean and variance of the weighted correlation of pitch contours was 0.78[13] and 0.04, respectively.

3.1 Pre-processing

Although we conducted practice trials, there still appears to have been a strong practice effect in our results. Figure 5 shows the average response time in chronological order. We can see a strong tendency toward decreasing response times as the experiment proceeds, especially at the beginning. Therefore, we excluded trials before the tenth trial from our results as training trials. This imbalance in the appearance of each morphing condition during preprocessing should be avoided in future research. Also, to control for individual differences in response times, we normalized the response times of each participant into z-scores for analysis as follows: $z = \frac{R-M_i}{\sigma_i}$ where R is the response time measured from the end of the speaker's production to the listener's keystroke response, M_i is the mean response time of participant i, and σ_i is the standard deviation of the response time of participant i.

3.2 Response time

3.2.1 Response time under different morphing conditions

Figure 6 shows the average normalized response times of each participant under different morphing conditions. Each color of bars show one participant's average response time under different morphing conditions. We can see that when participants heard voices the same or similar to their own (100% own voice and 67% own voice), they responded faster than when they heard voices dissimilar to their own (33% own voice and 0% own voice). But little difference was observed between the 100% own voice and 67% own voice conditions, or between the 33% own voice and 0% own voice conditions. Statistical analysis also supported this observation. Tukey's test indicates that there are significant differences between the 100% own voice and both the 33% own voice and 0% own voice levels ($p < 0.01$), and also between the 67% own voice and both the 33% and 0% own voice levels

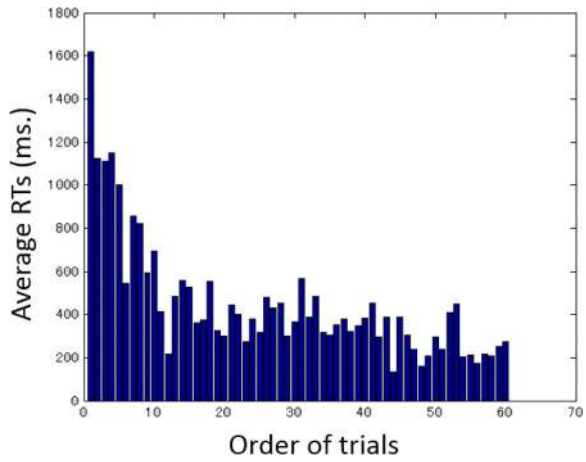


Fig. 5 Average response time for each trial. The horizontal axis represents the order of the trials, while the *vertical* axis represents the average response time of the *i*th trial from the end of the speaker's production to the listener's keystroke response

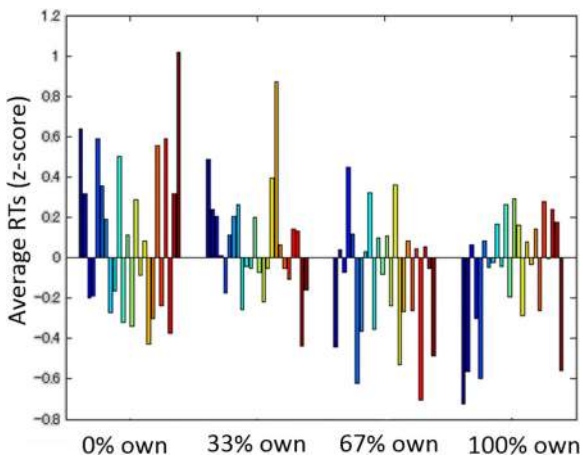


Fig. 6 Average response times of each participant under different voice morphing conditions. Each color of bars show one participant's average response time (z-score) under different morphing conditions

($p < 0.05$), but that there is no significant difference between the 100% own voice and 67% own voice levels, or between the 33% own voice and 0% own voice morphing conditions. It appeared that our participants could hardly distinguish the differences. We expected to find a linear relationship between morphing level and perceived similarity, but this was not the case. Thus, we combined the 100% own voice and the 67% own voice data and considered both to represent the "own voice" condition, while the 33% own voice and 0% own voice data were similarly combined to represent the "stranger's voice" condition.

Figure 7, shows a histogram of normalized response times for the "own voice" and "stranger's voice" conditions using the combination of morphing data percentages described above. Similar to the results shown in Fig. 6, participants responded faster when prosodic information was presented in voices similar to their own. The morphing conditions were considered as within-subjects factor (designs), statistical analysis (ANOVA) shows a significant difference between these two groups of normalized response times ($F = 15.22, p < .001$). As both of the participants in each pair experienced exactly the same stimuli (saw the same pictures and heard the same voices), we should be able to exclude the possibility of irrelevant factors, such as the match-up between the images and spoken words, that may cause a difference in response times. Therefore, the significant difference in response times is probably the result of the variation in the familiarity (similarity) of the voices presenting the information.

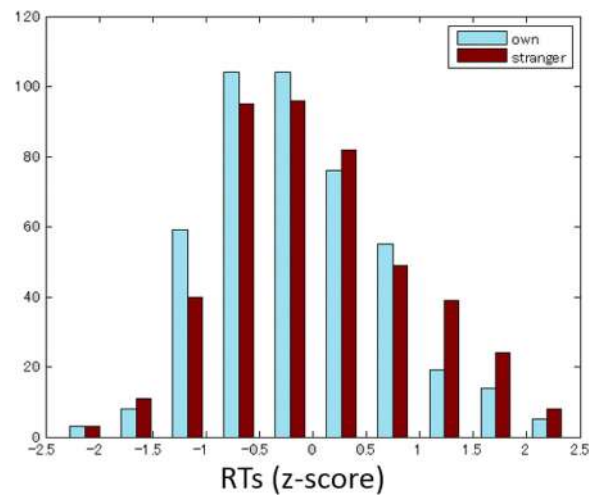


Fig. 7 Histogram of response times under different voice conditions. *Blue bars* stand for the "stranger's voice" condition (67% stranger's voice and 100% stranger's voice), and *red bars* stand for the listener's own voice condition (67% own voice and 100% own voice). *Horizontal axis* represents the normalized (z-score) response time

For example, in trial 1, partner A heard his own voice describing the objects, while partner B heard a stranger's voice (partner A) describing the objects in his experiment.

3.2.2 *Response time under different pairing conditions*

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

3.3 Degree of visual fixation

3.3.1 *Visual fixation under different voice morphing conditions*

We used an eye-tracking device to collect additional data to test our hypothesis. We analyzed the participants' degree of visual fixation on different areas of the target material in order to determine how much time they spent observing correct and incorrect images. The two rectangles, which contained the correct first item (no matter what the second item was) were defined as the "correct" areas, while the two rectangles, which contained the incorrect (ambiguous) first item were defined as the "incorrect" areas. Other parts of the screen, which had no items displayed were defined

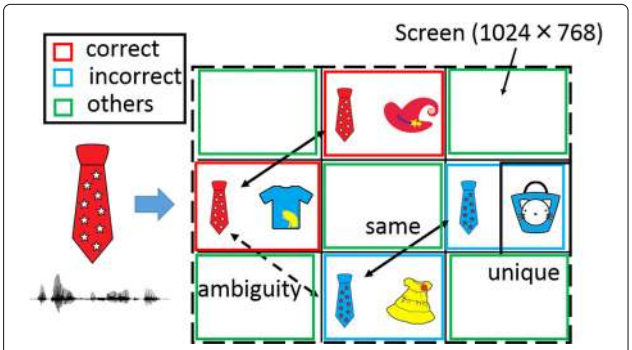


Fig. 8 Definition of visual fixation areas. When the first item is described ambiguously but with prosodic cues as "*red star necktie*," the areas inside the *red squares* are defined as "correct" areas, while the areas inside the *blue squares* are defined as "incorrect" areas. The other areas of the screen are defined as "other" areas

as "other" areas (Fig. 8). As explained in the "Procedure" section of this paper, in each trial the participants see four pairs of objects. The first item in each pair is described ambiguously, while the second item in each pair is described unambiguously. Until the description of the second item is provided, all of the rectangles containing the correct first item could be perceived by the participants as "correct" targets. In this experiment, we wanted to see whether there were differences in the proportion of visual fixation on "correct" areas under different voice morphing conditions. Figure 9 shows the proportion of listener eye fixation on the "correct" areas of the screen under different morphing conditions. We can see that although there is little difference during the second stage of the trial, participants were more likely to focus on the "correct" target during the third stage when the voice they were listening to was more similar to their own voice. The second stage includes the period from the beginning to the end of the description of the first item, and the third stage is listening to the word "and " followed by a short pause. These results support our hypothesis that listeners can more easily catch the subtle prosodic cues which help them to resolve lexical ambiguity when they are listening to voices similar to their own. There was no statistical difference between eye fixation on the "correct" and "incorrect" areas of the diagrams by the participants to confirm this, however.

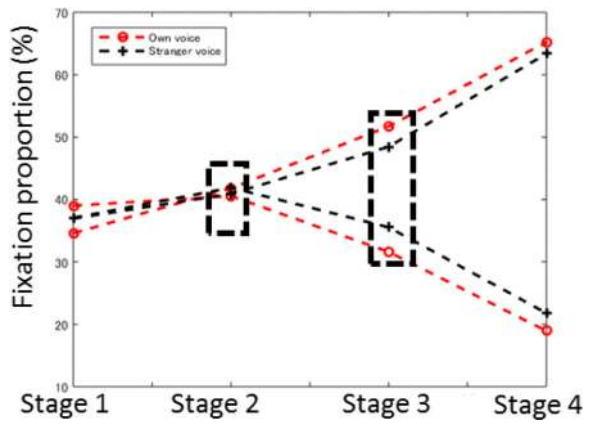


Fig. 9 Proportion of visual fixation on correct/incorrect areas under different morphing conditions during each stage of experimental trials. The *upper red line* shows the proportion of visual fixation on the area of the correct first item under the "own voice" condition (67% own voice and 100% own voice). The *lower red line* shows the proportion of visual fixation on areas of incorrect first items under the "own voice" condition. The *upper black line* shows the proportion of visual fixation on the area of the correct first item under the "stranger's voice" condition (67% stranger's voice and 100% stranger's voice). The *lower black line* shows the proportion of visual fixation on incorrect areas under the "stranger's voice" condition

3.3.2 Visual fixation under different pairing conditions

Just as in the previous section regarding response time under different pairing conditions, we further divided the "stranger's voice" condition into other voices similar to the listener's voice and other voices dissimilar to the listener's voice, and investigated differences in the visual fixation of the participants. Figure 10 shows the proportion of visual fixation on the "correct" areas under different spectrum similarity levels. Figure 11 shows the proportion of visual fixation on the "correct" areas under different pitch contour similarity levels (considered as within subjects factor). Figure 12 shows the proportion of visual fixation on the "correct" areas under different syllable/pause duration similarity levels (considered as within subjects factor)[14]. From these three figures we can see that when the listener hears another person's voice, which is similar to their own, their visual fixation during the third stage of the trials is the same as when they are listening to their own voice, especially when the trials are analyzed using spectrum and pitch contour similarity measurements. On the other hand, when listeners heard the voices of others, which differed from their own voices, we can see that their visual activity was more chaotic when selecting a fixation target. Statistical analysis shows a significant difference in the proportion of visual fixation on "correct" areas of the target when the "stranger's voices" were divided by spectrum similarity measure ($F = 4.64$, $p < 0.05$) and pitch contour similarity measure ($F = 8.32$, $p < 0.01$ as a between subjects factor, F=3.51, $p < 0.05$ as a within subjects factor). Also note that in Fig. 12,

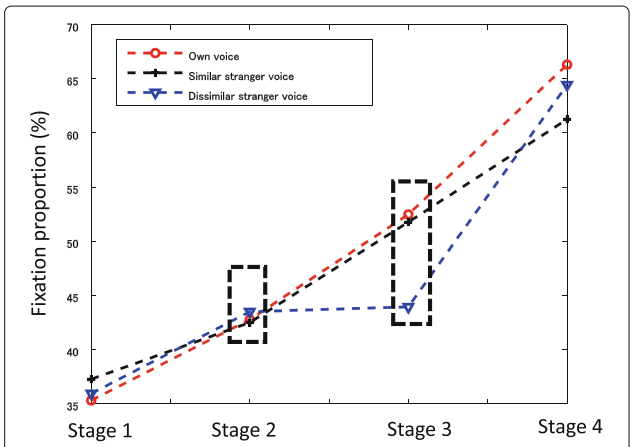


Fig. 10 Proportion of visual fixation on correct areas under different similarity conditions (DTW cost) during different trial stages. *Red* shows the proportion of visual fixation on areas with the correct first item under the "own voice" condition (same as in Fig. 9). *Black* shows the proportion of visual fixation on areas with the correct first item under the "similar stranger's voice" condition. *Blue* shows the proportion of visual fixation on areas with the correct first item under the "dissimilar stranger's voice" condition

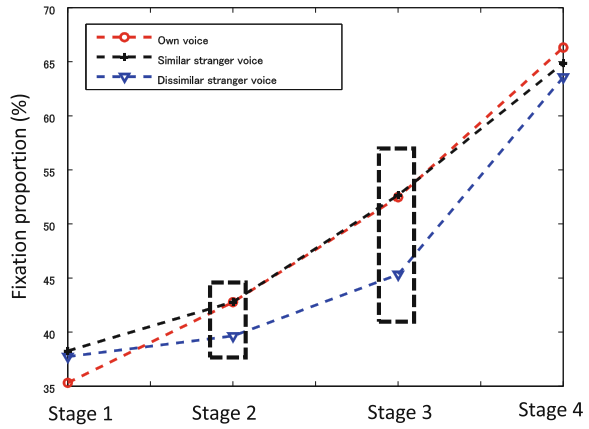


Fig. 11 Proportion of visual fixation on correct areas under different similarity conditions (pitch contour) during various trial stages. *Red* shows the proportion of visual fixation on areas with the correct first item under the "own voice" condition (same as in Fig. 9). *Black* shows the proportion of visual fixation on areas with the correct first item under the "similar stranger's voice" condition. *Blue* shows the proportion of visual fixation on areas with the correct first item under the "dissimilar stranger's voice" condition

while visual fixation on the "correct" areas under the "own voice" conditions and "similar stranger's voice" conditions are still similar, in contrast in Figs. 10 and 11, we can see that the proportion of visual fixation on "correct" areas is lower during the third stage under the "dissimilar stranger's voice" condition ($F = 0.36, p = 0.55$ as a between subjects factor, $F = 1.34, p = 0.27$ as a within subjects factor). This result may be because the duration cues used by different participants were perceptually more similar than the other two cues (i.e., changes in pitch and spectrum).

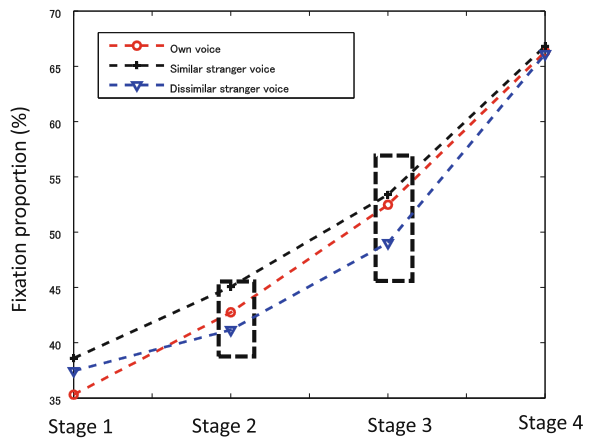


Fig. 12 Proportion of visual fixation on correct areas under different similarity conditions (duration) during various trial stages. *Red* shows the proportion of visual fixation on areas with the correct first item under the "own voice" condition (same as in Fig. 9). *Black* shows the proportion of visual fixation on areas with the correct first item under the "similar stranger's voice" condition. *Blue* shows the proportion of visual fixation on areas with the correct first item under the "dissimilar stranger's voice" condition

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

4 Conclusions

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

spectral envelope, influence efficiency of information transmission.

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

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

Endnotes

[1] The other ambiguous material we used can be found in the appendix.

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

[3] Considered to be the main prosodic cue.

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

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

[6] There are numerous ways to calculate the cost matrix, and here we only explain the method used in this paper (for more details see [20]).

[7] F_0 was tracked using TANDEM-STRAIGHT. Unvoiced intervals were interpolated based on a cost function aimed at minimizing discontinuities in the resulting trajectories and maximizing plausibility, based on the side information associated with F_0 candidates [21].

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

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

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

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

[12] We ignored participants/trials which did not meet both of the thresholds. For our analysis of spectrum similarity, we ignored two participants. For pitch similarity, we ignored three participants. For duration similarity, we ignored four participants.

[13] A value that has been considered to indicate a high level perceptual prosodic similarity in previous researches [19].

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

Appendix

Figure 13 shows the other 12 ambiguous materials we used in our experiment.

Fig. 13 The other 12 ambiguous items used in the experiment

Acknowledgements
This research is supported by the Center of Innovation Program (Nagoya-COI; Mobility Society leading to an Active and Joyful Life for Elderly) from Japan Science and Technology Agency.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate
This research has granted an approval from the ethics committee of the graduate school of information science, Nagoya University. The reference number of the ethics approval is 329. Participants were employed to attempt our experiment only after we received their e-mail informed consent.

Author details
[1]Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan. [2]Tokushima University, Minamijyosanjima-cho, Tokushima, Japan. [3]Institute of Innovation for Future Society (MIRAI), Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan. [4]JST/COI, Nagoya, Japan.

Received: 29 February 2016 Accepted: 15 November 2016
Published online: 13 December 2016

References

1. N Anderson, *Exploring Second Language Reading: Issues and Strategies.* (MA: Heinle and Heinle Publishers, Boston, 1999)

2. DI Slobin, Grammatical transformations and sentence comprehension in childhood and adulthood. J. Verbal Learn. Verbal Behav. **5**(3), 219–227 (1966)
3. L Frazier, K Rayner, Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. Cogn. Psychol. **14**(2), 178–210 (1982)
4. FC Bartlett, *Remembering: A Study in Experimental and Social Psychology*. (Cambridge University Press, Cambridge, 1995)
5. H Nassaji, Schema theory and knowledge based processes in second language reading comprehension: A need for alternative perspectives. Lang. Learn. **52**(2), 439–481 (2002)
6. MJ Leeser, Learner based factors in L2 reading comprehension and processing grammatical form: topic familiarity and working memory. Lang. Learn. **57**(2), 229–270 (2007)
7. SK Lee, Effects of textual enhancement and topic familiarity on korean EFL students' reading comprehension and learning of passive form. Lang. Learn. **57**(1), 87–118 (2007)
8. IH Erten, S Razi, The effects of cultural familiarity on reading comprehension. Read. Foreign Lang. **21**(1), 60–77 (2009)
9. P Adank, BG Evans, J Stuart-Smith, SK Scott, Comprehension of familiar and unfamiliar native accents under adverse listening conditions. J. Exp. Psychol. Hum. Percept. Perform. **35**(2), 520–529 (2009)
10. RC Major, SF Fitzmaurice, F Bunta, C Balasubramanian, The effects of nonnative accents on listening comprehension: implications for ESL assessment. TESOL Q. **36**(2), 173–190 (2002)
11. B Chen, N Kitaoka, K Takeda, *Relationship between speaker/listener similarity and information transmission quality in speech communication*. (Asia-Pacific Signal and Information Processing Association, Hong Kong, 2015), pp. 1190–1193
12. D Maurer, T Landis, Role of bone conduction in the self-perception of speech. Folia Phoniatr. Logop. **42**(5), 226–229 (1990)
13. Y Hirose, Cognitive mechanisms for sentence comprehension speaker's intention and hearer's comprehension: a latent function of lexical accent in syntax. Cogn. Sci. **13.3**, 428–442 (2006)
14. JJ Bartono, N Radcliffe, MV Cherkasova, J Edelman, JM Intriligator, Information processing during face recognition: the effects of familiarity, inversion, and morphing on scanning fixations. Perception. **35**, 1089–1105 (2006)
15. T Valentine, S Darling, M Donnelly, Why are average faces attractive? The effect of view and averageness on the attractiveness of female faces. Psychon. Bull. Rev. **11**(3), 482–487 (2004)
16. H Kawahara, T Takahashi, M Morise, H Banno, *Development of exploratory research tools based on TANDEM-STRAIGHT*. (Asia-Pacific Signal and Information Processing Association, Sapporo, 2009), pp. 111–120
17. VG Skuk, SR Schweinberger, Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. J. Speech Lang. Hearing Res. **57**(1), 285–296 (2014)
18. R Zaske, SR Schweinberger, H Kawahara, Voice aftereffects of adaptation to speaker identity. Hear. Res. **268**(1), 38–45 (2010)
19. DJ Hermes, Measuring the perceptual similarity of pitch contours. J Speech, Lang. Hear. Res. **41**(1), 73–82 (1998)
20. H Sakoe, S Chiba, Dynamic programming algorithm optimization for spoken word recognition. Acoust. Speech Signal Process. IEEE Trans. **26**(1), 43–49 (1978)
21. H Kawahara, A de Cheveigne, H Banno, T Takahashi, T Irino, *Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight*. (Interspeech, Lisbon, 2005), pp. 537–540