

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Evaluation of synthetic and natural Mandarin visual speech: Initial consonants, single vowels, and syllables

Trevor H. Chen*, Dominic W. Massaro

Department of Psychology, University of California, 1156 High Street, Santa Cruz, CA 95064, United States

Received 10 June 2010; received in revised form 8 February 2011; accepted 30 March 2011

Available online 7 April 2011

Abstract

Although the auditory aspects of Mandarin speech are relatively more heavily-researched and well-known in the field, this study addresses its visual aspects by examining the perception of both Mandarin natural and synthetic visual speech. In perceptual experiments, the synthetic visual speech of a computer-animated Mandarin talking head was evaluated and subsequently improved. Also, the basic (or “minimum”) units of Mandarin visual speech were determined for initial consonants and final single-vowels. Overall, the current study achieved solid improvements of synthetic visual speech, and this was one step towards a Mandarin synthetic talking head with realistic speech.

© 2011 Elsevier B.V. All rights reserved.

Keywords: Mandarin; Visual speech; Visemes; Speechreading; Synthetic; Natural; Talking head

1. Introduction

Visible speech synthesis can be valuable for scientific research and real-world applications (e.g., Caplier et al., 2007; Bailly et al., 2000; Cole et al., 1998; Massaro, 1998; Cohen and Massaro, 1990). An example is Baldi®, a computer-animated talking head (Massaro, 1998). This synthetic talking head has been used in experiments to test quantitative models of perception, such the fuzzy logical model of perception (FLMP; Massaro, 1998, 1987). This talking head has also been used to facilitate the learning of speech and language for hard-of-hearing children (e.g., Massaro and Light, 2004a,b), autistic children (e.g., Bosseler and Massaro, 2003), and learners of English as a second language (e.g., Massaro and Light, 2003). Of course, the value of visible speech is not limited to just the English language. Baldi has been adapted to speak a variety of languages, including Italian (Cosi et al., 2002), Arabic (Ouni et al., 2003), and Mandarin-Chinese (Massaro et al., 2006). Phonemes from the various languages are the basic

units of speech synthesis. The articulation properties of each phoneme are controlled by a set of facial animation control parameters (e.g., target values for jaw rotation, rounding, tongue position, etc.) and coarticulation temporal dominance functions (indicating the degrees of influence for their target values; Massaro et al., 2005). For a detailed description of the facial animation and speech synthesis involved in the talking head, see works by Massaro (1998, Chapter 12) and by Ouni et al. (2005, pp. 116–124).

Baldi speaks English, and its English visual speech has been evaluated and subsequently improved (Massaro, 1998, Chapter 13). Badr is the Arabic-speaking version of Baldi, and its Arabic visual speech has been evaluated (Ouni et al., 2003) and improved (Ouni et al., 2005). Bao is the Mandarin-speaking version of Baldi. Although Bao's Mandarin visual speech has been somewhat improved and used to train Mandarin learners (Massaro et al., 2006, 2008), its Mandarin visual speech repertoire has not been thoroughly evaluated and systematically improved. Therefore, the current experiments examine the visual speech perception of Mandarin natural speech and Bao's synthetic speech. The aim is to evaluate and improve Bao's Mandarin visual speech.

* Corresponding author. Tel.: +1 831 566 4249.

E-mail address: river_rover_t@yahoo.com (T.H. Chen).

Studying and applying Mandarin visual speech information potentially presents unique and interesting contributions. For example, Mandarin has phonemes that do not typically occur in English, including consonants (palatal fricative and palatal affricates) and vowels (high-front rounded and high-back un-rounded). Also, the Mandarin post-alveolar fricative, affricates, and approximant are apical (i.e., produced mainly with the tongue tip) instead of laminal (i.e., produced mainly with the tongue blade) (Lee and Zee, 2003). Moreover, Mandarin has some words consisting of syllabic consonants without true vowels (Lee and Zee, 2003), as in the pinyin words “zhi” and “zi”. Mandarin auditory speech is generally better understood than visual speech, and Mandarin offers an opportunity to study segmental and tonal aspects of visual speech.

Although Mandarin visual segmental speech has been examined from perspectives and approaches related to computer science and computer engineering (e.g., Pei and Zha, 2006, 2007), a search had failed to find psycholinguistic experiments published in English that specifically attempted to study the perception of Mandarin visemes. Visemes are units/categories of visual speech movements that are perceptually distinctive among the units/categories but much less so within each unit/category. The term “viseme” was used by Fisher (1968) for “visual phoneme”. A phoneme includes speech sounds that are perceived as the same auditory category, even though those sounds may have (slightly) different acoustic properties (e.g., the “s” in “sue” and “see” have different acoustics, but perceivers usually categorize them in the same /s/ category; Jackson, 1988). We use the term viseme, however, to simply describe a set of phonemes that have very similar visible properties and therefore are not easily distinguished from one another.

Understanding Mandarin visual speech from a psycholinguistic perspective, along with phonetic knowledge of Mandarin, would benefit science and computer engineering. In the field of synthetic visual speech, the chosen Mandarin visemes for consonants and vowels were at times noticeably inconsistent across studies (e.g., as a comparison among the following studies: Wu et al., 2006; Zhou and Wang, 2007; Chen et al., 2005; Ming et al., 1999). Perceptual–psycholinguistic experiments can refine our understanding of the functional visemes in Mandarin.

Wang et al. (2003) created a Mandarin text-to-visual speech synthesis system using a data-driven approach. To create Mandarin viseme categories, they analyzed certain static measurements (mouth measurements, jaw measurements, etc.) taken from video recordings of a Mandarin speaker. Based on the static parameters of the items (initials and finals), they constructed two confusion trees (one for consonant-initials, one for finals). The static parameters from an item are compared to those from every other item within the tree. The sum of the squared differences between the static parameters of the items is computed. The items that are closest in terms of the (normalized) sum of squared difference are treated as branches that merge into one category or trunk (Wang et al., 2003).

This method (Wang et al., 2003) is useful from a computer-science standpoint and informative to certain theoretical and applied perspectives. However, visemes also have a psycholinguistic dimension; production properties do not completely parallel perception. Just as differences in acoustic-physical features do not necessarily translate to differences in phonemes, differences in articulatory-physical measurements do not necessarily correspond to differences in visemes. In another article examining Mandarin speech-production (similar in approach to the one by Wang et al. (2003)), Wang et al. (2000) also acknowledge the importance of perceptual experiments for verification or refinement of visemes. In addition to visual-parameter analyses and Mandarin phonetics, the understanding and application of Mandarin visemes can benefit from perceptual-experimental studies using confusion matrices. Studies from multiple approaches can enrich our understanding. Most research regarding Mandarin visemes has been done from approaches in computer science and engineering. Mandarin visemes also need to be studied from a speech-science and psycholinguistic perspective, and the current experiments will add to our understanding of Mandarin visual speech as well as improving Bao’s visual speech repertoire.

The current experiments are not the first to address the perception of Mandarin visual segmental information. Two articles by Chen (1991, 1992), written in Chinese and published in Taiwan, examined Chinese (in Taiwan) college participants’ visual speech perception of Mandarin initial consonants (Chen, 1991) and vowel endings (single-vowels, diphthongs, and codas) (Chen, 1992). In these studies, the participants visually identified consonants and vowels by seeing the video of a female speaker pronouncing syllables that differed in initial consonants or differed in vowels (the video was without sound). All of the syllables were pronounced with tone 1. In the consonant study (Chen, 1991), all of the 21 Mandarin initial-consonants were followed by the single-vowel “a”, except for (pinyin) “j”, “q”, and “x”, which were followed by the diphthong “ia”. In the vowel study (Chen, 1992), a set of 16 endings, including single vowels, diphthongs, and codas, was presented without initial consonants.

On each trial, a stimulus was shown twice in succession (repetition) within the trial. Participants indicated their response on the response sheet, which displayed all of the possible answers to choose from. Each unique trial was presented only once in an experiment (for each of the participants; $n = 197$ in (Chen, 1991); $n = 200$ in (Chen, 1992)). To arrange phonemes into visemes based on the data, the author used agglomerative hierarchical cluster analysis with the method of average linkage between groups. A criterion was employed so that phonemes were in a viseme category that elicited at least 75%-correct responses on the viseme level. These two studies (Chen, 1991, 1992) present seminal, interesting, and informative data.

These earlier studies differ from the current experiments in terms of design, stimuli, data-collection, and data-

analysis. For example, in the studies by Chen (1991, 1992), there were two sets of stimuli presented separately that (mostly) differed only one part at a time: a set of 21 syllables that differed in initial consonants but are (mostly) kept constant in their vowel/ending (Chen, 1991), and another set of 16 vowels/endings without initial consonants (Chen, 1992). In the current experiments, the stimuli are 71 syllables that cover all initial consonants, single vowels (monophthongs), and their combinations. The current experiments randomly presented the 71 stimuli within the same experiment. Neither consonants nor vowels/endings were kept constant, and participants have to identify both consonants and vowels of the syllables/words together.

In terms of approach, it seemed that the studies by Chen (1991, 1992) were focused on the “basic” (or “ideal”) viseme categories that arise under conditions which minimize interactions between initial consonants and vowel endings. For the current experiments, the focus is examining viseme categories in the context of Mandarin words. This system of grouping will be used as a basis for assessing the performance of natural and synthetic speech, with the aim of improving the synthetic consonants, vowels, and whole syllables. In natural Mandarin characters, consonants and vowels necessarily influence and constrain each other within the Mandarin syllabic repertoire. Whether the performances and patterns of visemes are in a more “ideal/basic” setting or in a more “naturalistic” setting, both designs are informative and valuable. Table 1 lists and summarizes some of the differences. These differences may or may not lead to different viseme categories, and it is interesting to examine the viseme categories under these different environments.

2. Part I: Experiment A. Evaluation

Research examining speech-reading performance has applied various tasks at various linguistic levels, including

at the levels of syllable, word(s), phrase, sentence, story, etc. (Andersson et al., 2001; Mohammed et al., 2006). They can yield different performance accuracies. For example, Gailey (1987) examined speech-reading performance of English on a range of tasks. It was found that overall, tasks with non-sense syllables, words, and sentences (generally) yielded lower accuracies than tasks with stories, and tasks with stories yielded lower accuracies than tasks with familiar phrases (Gailey, 1987). On the other hand, this pattern of results was not always found. In the study by Mohammed et al. (2006), a speech-reading task with words yielded higher accuracies than tasks with minimal pairs and sentences, and all these types of tasks yielded higher accuracies than a task with short stories (connected speech of two to three sentences). Also, within each level, the specific tests and/or stimuli chosen can be different. The speech-reading abilities at different linguistic levels or tasks may require different skill sets (e.g., Andersson et al., 2001), and they are not necessarily related to each other (Gailey, 1987), although even this aspect has been disputed, adding to the complexity of the situation.

What is clear is that, to date, most of the research on speech-reading has been carried out on “European/Western” languages (especially English). The present experiments examine the visual speech-reading of Mandarin-Chinese. The focus is at the character-syllable level of analysis, which may be a more “pure” and “perceptual” form of speech-reading (as opposed to, for example, the sentence level), as this level may greatly reduce (or minimize) some possible conceptual or contextual influences. At higher linguistic levels, syntactic, semantic, and pragmatic contexts can influence the accuracy of speech-reading performance. From a practical perspective, Mandarin individual characters are essentially single syllables. In order to compare and evaluate the perception of natural and synthetic speech, natural visual speech has to be recorded. Given the Mandarin syllabic-lexical structure, all syllables

Table 1
Some differences between the current experiments and studies by Chen (1991, 1992).

	Chen (1991, 1992)	The current experiments
Stimuli	<ul style="list-style-type: none"> – 21 Syllables that differed in initial consonants – 16 Vowels/endings without initial consonants – 2 Sets presented separately – Some are non-words 	<ul style="list-style-type: none"> – 71 Syllables covering all initial consonants, single vowels, & their combinations – Without codas & diphthongs – Presented together randomly – All are Mandarin words
Natural speaker pronunciation	“Non-exaggerate” manner	Told to pronounce “clearly”
Design	<ul style="list-style-type: none"> – A lot of participants – Each unique trial presented once to a participant 	<ul style="list-style-type: none"> – Fewer participants – Each unique trial presented 10 times (10 blocks) for every participant
Data-collection	“Close-ended”: the possible choices on the response sheet	Open-ended: could type any Mandarin word/syllable
Viseme analysis	<ul style="list-style-type: none"> – Agglomerative hierarchical clustering analysis – Criterion of (at least) 75% correct on the viseme level 	<ul style="list-style-type: none"> – Clustering analysis with correlation (response patterns) as the metric – d-Prime analyses – Mandarin phonetics

within Mandarin speech are characters. Using the Mandarin syllable level allows practical analyses of the phoneme-viseme, character/word, and tonal domains.

The present experiments examine and compare the perception of natural and synthetic Mandarin visual speech. The goals are to analyze the performance of phonemes, syllables/words, and viseme categories of all Mandarin syllables involving initial consonants and single-vowel endings, examine the perceptual performances of natural and synthetic visual speech, determine where the synthetic visual speech under-performs their natural counterparts, and subsequently improve and re-evaluate the visual speech information.

2.1. Method

2.1.1. Participants

In this 4-h experiment, four Chinese participants were recruited from the University of California, Santa Cruz (UCSC). They are all native speakers of Mandarin and are all familiar with pinyin. They consist of two females and two males, and they were at the ages of 21, 20, 22, and 21. Their lengths of time living in the US had been 13, 13, 10, and 11 years. Their exposure to English began at the ages of 8, 7, 12, and 10. They all had reported to have never taken a course in linguistics or phonetics. For their participation, they either received course credits, payment at the rate of \$10 per hour, or a combination of course credit(s) and payment.

2.1.2. Stimuli

For natural-speech stimuli, we made sets of audio–video recordings consisting of words pronounced by four speakers: One female from Mainland China (FC), one male from Mainland China (MC), one female from Taiwan (FT), and one male from Taiwan (MT). The words cover all possible Mandarin segments. Speakers from Mainland China read abbreviated (simplified) characters, and speakers from Taiwan read traditional characters. The words to be read (in citation form) were randomized and displayed as slides (Microsoft Powerpoint™) on a standard computer screen, and there were approximately five seconds between the word presentations. The speakers were told to pronounce the characters (single-words) clearly. The camera was right next to the computer screen (on its immediate right), capturing a clear frontal view of the speaker (albeit slightly angled to her/his left). We used an ordinary camera and tripod. There was lighting at the front (front-right and front-left) of the speaker and above the speaker. It was under similar conditions as the audio and video recordings from the study by [Chen and Massaro \(2008\)](#). The speech recordings were transformed into computer AVI files. For this experiment, stimuli from FC were used, because it was observed that her speech was the most visually clear, and a previous experiment found that her visual speech yielded higher performance than the visual speech of another speaker did. Also, she was taught “standard Mandarin”,

and she did not speak another Chinese dialect (Mandarin was the only Chinese language she spoke). The synthetic-speech stimuli were the same Mandarin words articulated by Bao – that is, a computer-animated female version of Bao (with dark hair and dark eyes). Each of the synthetic words was aligned to the corresponding natural word to ensure that the durations would be constant – this was to control for a possible effect of duration on speech-reading performance.

2.1.3. Procedure

For this experiment, there were 71 word-syllables (which cover all Mandarin initial consonants, all Mandarin single-vowel endings, and all of their combinations) from each of the 2 speakers (synthetic and natural), yielding 142 unique stimuli ($71 \times 2 = 142$) that were randomized and presented within each of the 10 blocks. Thus, the total number of trials was $[(71 \text{ syllables}) \times (2 \text{ speakers}) \times (10 \text{ blocks}) = 1420]$ 1420 for each participant. The 71 word-syllables are listed in [Appendix A](#). Out of the 71 syllables, there were 39 tone-1 syllables, 12 tone-2 syllables, 6 tone-3 syllables, 13 tone-4 syllables, and 1 tone-0/5 syllable. For each trial, the stimulus video was preceded by 750 ms (ms), and each video included about 500 ms before and 500 ms after the actual articulation (the natural or synthetic face remained on the screen during those 500 ms times). The participants were instructed to watch the video (with no sound available) on the computer screen and type the word that they thought was spoken. They were instructed to type in pinyin-words with a space between the pinyin and the numerical tone (1, 2, 3, 4; the light/neutral tone was typed as either 0 or 5, depending on their personal preference). No feedback was given.

Because whatever the participants typed would be showing on a space near the bottom of the computer screen until the next trial appears, they could use the “Backspace”, “Delete”, arrow keys, or other buttons to modify what they typed for a given trial (in case there was a typo or mistake) before they press the “Enter” key. The experimental trials were response-paced: After the participant types a response and presses “Enter”, then the typed-response for the trial would disappear on the screen and the next stimulus would appear. There was roughly 5–10 min of break between blocks. The participants scheduled for two-hour and/or one-hour sessions (usually on different days) until they completed the whole experiment, which usually took about 4 h total. All instructions and interactions were spoken in Mandarin.

2.2. Results

Before the responses were grouped into viseme categories, an “absolute performance” was first calculated. An absolute performance is indicated by the percentage of correctly-identified syllables, even if the mistake is within a viseme category (the tone did not have to be correct). This performance measure offers an estimate of “real-world”

performance and an additional estimate of the relationship between the perception of natural and synthetic speech. In terms of “absolute performance”, natural speech yielded 11.58% average correct segments, and synthetic speech yielded 5.92% average correct. A correlation was carried out between natural speech and synthetic speech on number of correct responses for the 71 stimuli (syllables). There was a significant correlation between the response patterns of the natural syllables and those of the synthetic syllables [$r(69) = .72$, $p < .001$] – an indication that performance under the natural-speaker condition was somewhat similar to performance under the synthetic-speaker condition.

Given the fuzzy nature of visual speech (Massaro, 1998, p. 395), there is a lack of agreement regarding how visemes should be analyzed, and there is no universally recognized rule or criterion for grouping phonemes into visemes (Jackson, 1988). Various criteria or approaches can be employed that render somewhat-different or slightly-different categories (for English cf., Campbell and Massaro, 1997; Massaro, 1998, pp. 412–413; Jackson, 1988). For the current study, several approaches are considered in combination to determine the Mandarin viseme categories. Overall, the viseme groupings should: (A) be consistent with what is known about Mandarin phonetics (e.g., Lee and Zee, 2003) in particular and phonetic theory (e.g., Pullum and Ladusaw, 1996; Ladefoged, 2001) in general (unless the data overwhelmingly suggest otherwise); (B) include phonemes that render similar response-patterns (significant and high-positive correlations between the responses-patterns of the phonemes); (C) be in general agreement with other perceptual-experimental research findings; (D) yield the highest d' -prime (d' – a measure of how well each item is distinguished from the others) values (for an explanation on d' , see Massaro, 1989a; Macmillan and Creelman, 2005) when grouped this way; and (E) have no apparent reason of why they should be grouped otherwise. In other words, the behavioral data are considered along with linguistic theory, other relevant findings, and an analytical approach that practically group phonemes together based on the similarity of their response-patterns (i.e., items in a group are the closest to each other, more similar to each other than to those outside the group).

One way to see that two phonemes are confused together is by looking at the degree to which their

response-patterns (generated by the confusion matrix) are similar to each other. Similarity of response patterns can be reflected by the pattern of mutual confusion, but it can also include other arrangements. A good example is the Mandarin “b”, “p”, and “m”. These three phonemes were confused with each other but almost never confused with any other phoneme. When they were grouped together in a viseme category, 99% of the responses are within this category (Fig. 1). Other phonemes were rarely categorized as “b”, “p”, or “m”. They are all bilabials (fits with criterion A); their response-patterns were similar to each other and not to those of other phonemes – their correlations were highly-positive and significant [their r 's(20) $> .99$, p 's $< .001$] (fits with B); the grouping is consistent with other research (e.g., Walden et al., 1977) (fits with C); they yielded the highest d' when grouped this way (fits with D); and there seemed to be no apparent reason regarding why they should be grouped otherwise (fits with E).

Although these three phonemes were highly confused with one another, their response-patterns were not symmetrical. All three phonemes yielded mostly “b” responses, followed by “p” and then “m” responses. Because the response-patterns were highly similar among these three phonemes, they are confused together. To the extent that phonemes yield similar response patterns across the board, and that it makes linguistic-phonetic sense, one can assume that participants do not distinguish much between them.

Based on the current combination of criteria (and clustering analyses), the viseme categories are grouped as follows. The categories, expressed in pinyin, are: “b/p/m”, “f”, “d/t/n/l”, “z/c/s”, “zh/ch/sh/r”, “j/q/x”, “g/k”, and “h”. Not surprisingly, all phonemes within the viseme groups have the same place of articulation. The category “d/t/n/l” was not immediately obvious, given that Zhou and Wang (2007) put “d/t/n/l” in the same viseme category, while Wu et al. (2006) put “l” in a separate viseme from “d/t/n”. From examining the experimental stimuli, it was discovered that the natural Mandarin speech involving initial consonants “d”, “t”, “n”, and “l” were articulated with the “dental-alveolar” tongue clearly visible. While in English, “d”, “t”, “n”, and “l” are usually considered alveolar (e.g., Ladefoged, 2001), their Mandarin counterparts can be different. In Mandarin, “d”, “t”, “n”, and “l” can be categorized as dental-alveolar for their place

Stimuli.C	B/P/M	F	D/T/N/L	Z/C/S	ZH/CH/SH/R	J/Q/X	G/K	H	none	Total
B/P/M (100)	99.00	0.50	0.00	0.00	0.00	0.00	0.00	0.25	0.25	100
F (20)	1.25	97.50	0.00	0.00	0.00	0.00	0.00	1.25	0.00	100
D/T/N/L (180)	0.00	0.14	70.56	0.56	9.17	1.94	0.69	4.58	12.36	100
Z/C/S (90)	0.00	0.00	20.56	21.11	28.89	21.94	0.28	1.39	5.83	100
ZH/CH/SH/R (110)	0.00	0.00	11.82	10.00	37.73	29.77	0.00	0.68	10.00	100
J/Q/X (60)	0.00	0.00	14.17	7.08	25.42	24.58	0.00	0.00	28.75	100
G/K (60)	0.42	0.00	45.00	4.58	3.33	6.67	0.83	17.92	21.25	100
H (30)	0.83	0.00	10.83	0.00	0.83	0.00	0.83	58.33	28.33	100
none (60)	0.00	0.00	21.25	0.42	5.83	2.08	0.00	28.75	41.67	100
Grand Total	101.50	98.14	194.18	43.75	111.20	86.99	2.64	113.15	148.44	900

Fig. 1. The percentage of responses for the natural initial-consonant visemes.

of articulation (e.g., Lee and Zee, 2003). Mandarin dental-alveolar consonants are not confused with inter-dentals, because Mandarin does not have inter-dentals. The d' calculations revealed that, whether “l” was included in “d/t/n” or not, the d' values were about the same (1.24), so for the sake of parsimony and partly based on criterion E, “l” was grouped with “d/t/n”. This grouping was further supported by the finding that the response-patterns between “d”, “t”, “n”, and “l” were all similar to each other (they were the closest to each other); the correlations between the response-patterns of each of these phonemes were all highly-positive and significant [r 's(20) > .90, p 's < .001].

Observations of Mandarin speakers in natural settings suggested that some Mandarin speakers tend to articulate “d/t/n/l” with dental or dental-alveolar tongue, some tend to articulate them with alveolar tongue, and yet others may articulate them in both ways (in different contexts). Phonemes d/t/n can be dental or alveolar (Pullum and Ladusaw, 1996). Presumably, when their place of articulation is dental or dental-alveolar (as in the current experimental natural-speech stimuli), they seemed visually informative in Mandarin. When their place of articulation is mostly alveolar, they may be considered less visually-informative (relative to some of the other visual consonants), at least in English (e.g., Jackson, 1988). It turns out that, when this experiment's data are combined/pooled with the next experiment's data, a clustering analysis (using

correlation as the metric) further supports each of the current grouping-arrangement of visemes (more on that appears in the next part).

Fig. 1 shows the average percentage of responses for the natural initial-consonant visemes. Fig. 2 shows the responses for the synthetic counterparts. The “none” labels refer to the stimuli or responses that have no initial consonants.

For single-vowel endings: Fig. 3 shows the average percentage of responses for the natural single-vowels, and Fig. 4 shows the responses for the synthetic counterparts. As with the consonants, the labels for the vowels are written in pinyin spelling. The response-patterns seemed similar between “a” and “er” [$r(7) = .99$, $p < .001$] and between “u” and “v” (“v” in pinyin denotes the front-high rounded vowel) [$r(7) = .91$, $p = .001$]. When the category “other” (diphthongs and triphthongs) was divided to show those responses that included the target vowels, the response-patterns were still similar between “a” and “er” and between “u” and “v”. An analysis on the d' values showed that when “u” and “v” were grouped together, the d' value was higher ($d' = 2.33$) than when “u” and “v” were grouped as separate visemes ($d' = 1.91$). The grouping of “u” and “v” together made sense because “u” is a high-back rounded vowel, and “v” is a high-front rounded vowel (fit criterion A); their response-patterns were similar (the correlation was highly-positive and significant between

Stimuli.C	B/P/M	F	D/T/N/L	Z/C/S	ZH/CH/SH/R	J/Q/X	G/K	H	none	blank/error	Total
B/P/M (100)	97.25	1.50	0.00	0.25	0.00	0.25	0.00	0.25	0.50	0.00	100
F (20)	3.75	92.50	0.00	0.00	0.00	2.50	0.00	1.25	0.00	0.00	100
D/T/N/L (180)	0.97	1.53	23.06	14.31	16.39	19.58	0.83	4.44	18.89	0.00	100
Z/C/S (90)	1.11	3.61	8.61	22.50	24.44	24.17	0.83	2.22	11.94	0.56	100
ZH/CH/SH/R (110)	0.68	0.45	12.27	13.64	34.77	25.00	1.36	2.27	9.55	0.00	100
J/Q/X (60)	0.83	2.92	14.17	5.42	21.25	24.17	0.83	2.92	27.50	0.00	100
G/K (60)	0.42	0.42	35.83	10.83	21.67	11.67	0.42	2.50	16.25	0.00	100
H (30)	0.83	0.00	14.17	11.67	21.67	25.83	0.83	2.50	22.50	0.00	100
none (60)	2.50	9.17	20.42	10.42	7.50	11.25	0.42	6.67	31.67	0.00	100
Grand Total	108.35	112.09	128.52	89.03	147.69	144.42	5.53	25.02	138.80	0.56	900

Fig. 2. The percentage of responses for the synthetic initial-consonant visemes.

Stimuli.V	a (an, ang)	e	er	i (in, ing)	u.v (ong, vn)	none (0)	en (eng)	Dip(n).Trip	error/blank	Total
a (180)	26.25	0.00	0.00	0.00	0.00	0.00	0.00	73.61	0.14	100
e (160)	0.47	36.25	0.94	21.41	1.09	3.13	7.34	29.38	0.00	100
er (10)	32.50	2.50	0.00	0.00	0.00	0.00	0.00	65.00	0.00	100
i (110)	0.00	2.50	0.00	86.82	0.00	5.91	0.00	4.77	0.00	100
u.v (250)	0.00	0.00	0.00	0.00	95.90	0.20	0.00	3.50	0.40	100
Grand Total	59.22	41.25	0.94	108.22	96.99	9.23	7.34	176.26	0.54	500

Fig. 3. The percentage of responses for the natural single-vowel endings.

Stimuli.V	a (an, ang)	e	er	i (in, ing)	u.v (ong, vn)	none (0)	en (eng)	Dip(n).Trip	error/blank	Total
a (180)	18.89	0.69	0.00	0.28	0.14	0.14	0.56	79.17	0.14	100
e (160)	12.66	12.50	0.00	0.31	0.63	0.00	2.34	70.94	0.63	100
er (10)	0.00	7.50	0.00	47.50	2.50	12.50	0.00	30.00	0.00	100
i (110)	1.36	3.41	0.00	48.18	4.32	10.91	0.23	30.91	0.68	100
u.v (250)	0.70	3.00	0.10	0.70	57.90	0.60	0.10	35.50	1.40	100
Grand Total	33.61	27.10	0.10	96.97	65.48	24.15	3.23	246.51	2.85	500

Fig. 4. The percentage of responses for the synthetic single-vowel endings.

their response-patterns and not between theirs and those of other phonemes) (fit B); it is generally not inconsistent with other findings (fit C); the arrangement rendered higher measure of d' (fit D); and there seemed to be no reason not to do so (fit E).

While the grouping of “u” and “v” together fit the current criteria, “a” and “er” were not grouped together for now; it was not clear whether grouping “a” and “er” together would be consistent with phonetics. Also, observations of the natural stimuli suggested that, while “u” and “v” appeared to look “indistinguishable”, “a” and “er” did appear to look somewhat “different”. Another caution was that “er” consisted of only one stimulus out of the 71 stimuli, and it does not appear with anything else in a “standard” Mandarin syllable (unlike other single vowels), except for some regional linguistic variations: In Mainland China, this is generally true in “Southern” Mandarin speech, but this is not the case in some “Northern” Mandarin speech. In Taiwan, this is typically true. Given the special status of “er”, it seemed “safer” to examine it by itself for now. As it later turns out, when a clustering analysis combines/pools data from this experiment and data from the next experiment, the correlation between “a” and “er” would be much lower and non-significant, and the correlation between “u” and “v” would be the only highly-positive and statistically-significant correlation among the single-vowel endings.

2.3. Discussion

In this experiment, although there were only four participants, the response patterns were highly similar among all four of the participants. For the initial consonants: All correlations among the participants for the response patterns of the phonemes were significant [r 's(482) ranged from 0.65 to 0.80, all p 's < .001]. When the response patterns were grouped into the viseme categories, all of the corresponding correlations among participants were again significant but even greater than the previous coefficients [r 's (79) ranged from 0.83 to 0.96, all p 's < .001]; these correlation coefficients calculated from data based on the viseme categories were significantly higher than those based on the phoneme categories [$t(5) = 3.91$, $p < .05$], with the average correlation coefficient being 0.14 higher.

As for the single-vowel endings: All correlations among the participants' response patterns were significant [r 's(52) ranged from 0.72 to 0.92, all p 's < .001]. When “u” and “v” (the two high-rounded single-vowels in pinyin) were grouped into a category, all of the corresponding correlations among participants were also significant but even greater [r 's(38) ranged from 0.86 to 0.97, all p 's < .001]; the correlation coefficients calculated with merging “u” and “v” were significantly higher than those calculated from the phoneme categories [$t(5) = 2.91$, $p < .05$], with the average correlation coefficient being 0.09 higher.

Generally, the Mandarin synthetic initial-consonants from Bao yielded lower percents of correct compared to

the natural initial-consonants. Although performance from the synthetic labials and labio-dental were comparable to the natural ones, the synthetic non-labials need specific improvements. Synthetic “d/t/n/l” and “h” especially need improvement, as they yielded performances that were about 50% lower than those from their natural counterparts. As for the single-vowel endings: Synthetic single-vowels from Bao yielded lower percents of correct compared to the natural ones, especially for “e” (a Mandarin-unique, mid-high back, un-rounded vowel), “i”, and “u/v”. The subsequent experiment aims to modify/improve the synthetic initial-consonants and single-vowel endings and re-evaluate the visual perception of them along with their natural counterparts.

3. Part II: Experiment B. Improvement

In the previous experiment, Bao's synthetic initial-consonants and single-vowels yielded performance that was less accurate than performance yielded by their natural counterparts. For this subsequent experiment, a set of improved synthetic stimuli are tested with the same natural stimuli from the previous experiment. Comparisons of performance yielded by the natural stimuli and the synthetic stimuli in both experiments will assess whether Bao's synthetic speech is improved (i.e., more “realistic”) from the previous experiment to this experiment. In addition, data yielded by natural stimuli can be combined/pooled across the two experiments, and these data can be analyzed to help understand the perception of Mandarin visemes in the context of perceiving words/syllables.

3.1. Method

3.1.1. Participants

In this 4-h experiment, ten Chinese participants were recruited from UCSC. They are all native speakers of Mandarin and are all familiar with pinyin. They consist of five females and five males, and their average age was 25.50 years old. Their average length of time living in the US had been 1.50 years, and their exposure to English began at the average age of 9.70. They all had reported to have never taken a course in linguistics or phonetics. For their participation, they either received course credits, payment at the rate of \$10 per hour, or a combination of course credit(s) and payment.

3.1.2. Stimuli and procedure

The natural-speech stimuli were the same visual-speech stimuli used in the previous experiment (part I). On the other hand, the synthetic-speech stimuli were the matching Mandarin words/syllables articulated by Bao (again, the computer-animated female version of Bao), but these stimuli were specifically improved by modifying Bao's parameters. Each of the individual synthetic words was directly compared to its natural counterpart. The parameters of Bao's consonants and vowels were adjusted until they were

deemed authentic/realistic. Modifications included changing the “mouth shapes” or articulations of the phonemes, making the tongues in “d”, “t”, “n”, and “l” look more like each of its natural counterpart, increasing the vertical mouth opening in “a”, and many more. Table 2 lists and summarizes the main differences of the stimuli from the previous experiment to the current experiment. All of the 71 individual words/syllables were made to look like each of its natural counterpart. After all of the stimuli were individually prepared, they were tested for experimental verification. Like the previous experiment, each of the synthetic words was aligned to the corresponding natural word (even though no sound was present in the test stimulus) to ensure that the durations would be constant. The procedure in this experiment is the same as the previous experiment (part I). All instructions and interactions were spoken in Mandarin.

3.2. Results

In terms of “absolute performance”, natural speech yielded 11.38% average correct segments, and synthetic speech yielded 6.54% average correct. Comparing to the previous experiment, the natural–synthetic gap reduced from 5.66 to 4.84 (a 14.49% reduction). The correlation between natural and synthetic speech on the number of correct syllables was significant [$r(69) = .79, p < .001$], indicating that performance under the natural-speaker condition is somewhat similar to performance under the synthetic-speaker condition.

3.2.1. Initial consonants

Fig. 5 shows the average percentage of responses for the natural initial-consonant visemes. Fig. 6 shows the responses for the synthetic counterparts.

In terms of the overall percentage of correct responses (viseme-based), this experiment’s natural speech yielded 50.14% correct responses, and its synthetic speech yielded 42.44% correct responses, so the natural-minus-synthetic overall difference is 7.70%. Comparing to the previous experiment, in which the natural-minus-synthetic gap was ($47.71\% - 32.57\% = 15.14\%$) 15.14%, this natural–synthetic difference reduced by a factor of 49.14%, almost cutting the advantage of the natural speech in half. Synthetic viseme categories that previously yielded the worst performances now showed the biggest improvements. Specifically, in the previous experiment, the natural–synthetic gap for d/t/n/l was 47.50%, and it was 55.83% for h. Those were the two biggest gaps. Now, in the current experiment, these gaps are now 8.33% and 17.33%, respectively.

Also, d' values were calculated for each of the viseme categories. In terms of the overall d' values, the previous experiment yielded 1.89 for its natural visemes and 1.05 for its synthetic visemes (a natural–synthetic gap of 0.84). The current experiment yielded 2.07 for its natural visemes and 1.37 for its synthetic visemes (a natural–synthetic gap of 0.70). Synthetic visemes from the current experiment have larger d' values than synthetic visemes from the previ-

ous experiment, and the natural-minus-synthetic gap for d' was reduced by a factor of 16.67%. Fig. 7 shows the natural–synthetic gaps of the d' values for both experiments. All of the d' gaps from the current experiment (experiment 2) are either smaller than or comparable to those from the previous experiment (experiment 1), except for viseme z/c/s. The d' difference for “f” is smaller than it appears: In Experiment 2, the natural “f” had a 100% hit rate and a 0% false-alarm (FA) rate. In this case, the d' would have to be estimated, and it was calculated from a 99.50% hit rate and 0.01% FA rate (closest possible rates in this condition), producing a large d' value for the natural “f” and increasing the natural–synthetic gap for “f” (Experiment 2). If only the d' values of the synthetic speech are compared across experiments, then all of the d' values of synthetic visemes from the current experiment are bigger than those from the previous experiment, except for viseme z/c/s.

Another comparison was the correlation between the response pattern (from confusion matrices) yielded by the natural speech and that yielded by the synthetic speech. In terms of this correlation, the current experiment’s correlation [$r(62) = 0.97, p < .001$] was higher than that of the previous experiment [$r(62) = 0.81, p < .001$]. Putting the three measures (% correct, d' , and r) together, the results suggested that, comparing synthetic speech in the current experiment to that in the previous experiment, the current set of synthetic speech produced responses that were more accurate, and its response patterns are more similar to those of the natural speech.

3.2.2. Single-vowel endings


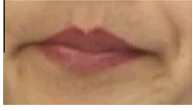


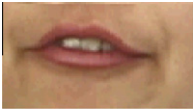

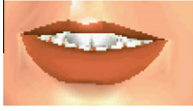
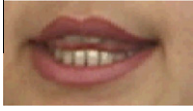


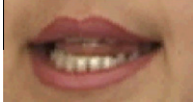


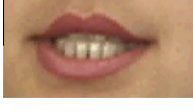
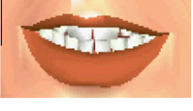
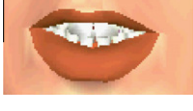
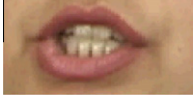
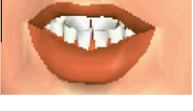
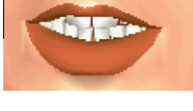
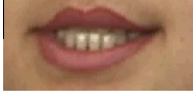


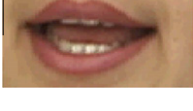

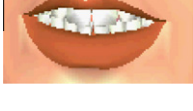
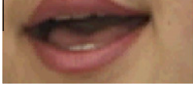


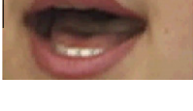
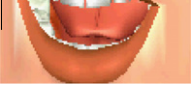
Fig. 8 shows the average percentage of responses for the natural single-vowel endings, and Fig. 9 shows the responses for the synthetic counterparts.

In terms of the overall percentage of correct responses, this experiment’s natural speech yielded 57.68% correct responses, and its synthetic speech yielded 47.31% correct responses, so the natural-minus-synthetic overall difference is 10.37%. Comparing to the previous experiment, in which the natural-minus-synthetic gap was ($62.04\% - 35.46\% = 26.58\%$) 26.58%, the natural–synthetic difference reduced by a factor of 60.99%, cutting the overall advantage of the natural speech by more than half.

All synthetic vowels in this experiment showed improvement compared to the previous experiment, except for “er”. Specifically, for Mandarin vowel “a”, the previous natural–synthetic gap of 7.36% is now almost eliminated. Although it appeared as if the current synthetic “a” outperformed the natural “a”, this is not really the case. Natural stimuli with “a” yielded a majority of responses as diphthongs or triphthongs involving “a” as a constituent; “a” was the only natural vowel-ending that produced more responses in diphthongs or triphthongs than “correct” responses as a single-vowel (besides “er”). So, if diphthongs or triphthongs involving “a” are also counted as correct responses, then the natural–synthetic gap for “a” is


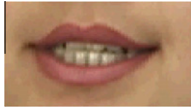


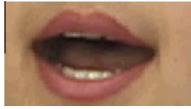


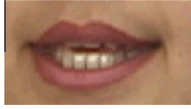

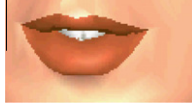
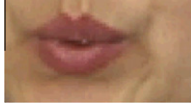
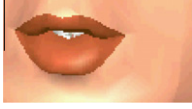
Table 2

Summary of the main changes in synthetic stimuli between experiment 1 and experiment 2.

“Viseme” pinyin, phonetic description	Main changes	Experiment 1: synthetic (representative)	Natural speech (representative)	Experiment 2: synthetic (representative)
b/p: Bilabial plosive m: bilabial nasal	Not much; making sure that they look ok in all cases			
f: Labio-dental fricative	Not much; making sure that they look ok in all cases			
d/t: Dental-alveolar plosive n: dental-alveolar nasal	Dental-alveolar tongues; more mouth opening; somewhat different depending on vowels			
l: Dental-alveolar lateral approximant				
z/c: Dental-alveolar affricate; s: dental-alveolar fricative	More mouth opening; teeth in certain position			
zh/ch: Post-alveolar affricate (apical) sh: post-alveolar fricative (apical); r: post-alveolar approximant (apical)	More refined mouth shape; a bit more rounding; teeth position			
j/q: Palatal affricate; x: palatal fricative	Mouth shape; showing a bit more lower teeth			
g/k: Velar plosive	Mouth opening and shape; vowel dependent			
h: Velar fricative	More mouth opening; tongue; vowel dependent			
a: Low mid-front un-rounded	Jaw rotation/mouth opening			

(continued on next page)

Table 2 (continued)

“Viseme” pinyin, phonetic description	Main changes	Experiment 1: synthetic (representative)	Natural speech (representative)	Experiment 2: synthetic (representative)
e: High-back un-rounded (ram’s horns)	Teeth position; mouth shape; less opening			
er: Mid-central schwa with curly-r tail	Mouth opening; tongue			
i: High-front un-rounded	A bit more mouth opening; teeth; spread			
u: High-back rounded; v: high-front rounded	More rounding; associated facial contours			

Stimuli.C	B/P/M	F	D/T/N/L	Z/C/S	ZH/CH/SH/R	J/Q/X	G/K	H	none	blank/error	Total
B/P/M	99.00	0.00	0.20	0.00	0.00	0.10	0.10	0.00	0.40	0.20	100
F	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100
D/T/N/L	0.17	0.00	67.33	2.22	8.61	6.50	1.44	1.33	12.28	0.11	100
Z/C/S	0.00	0.00	8.44	25.00	41.78	16.89	2.00	0.33	5.56	0.00	100
ZH/CH/SH/R	0.00	0.00	5.36	9.73	58.91	19.09	0.18	0.27	6.36	0.09	100
J/Q/X	0.00	0.00	8.17	8.00	36.17	30.50	0.33	0.17	16.33	0.33	100
G/K	0.17	0.00	39.83	2.33	9.17	8.50	5.00	7.33	27.33	0.33	100
H	0.00	0.00	13.67	0.33	3.33	2.67	3.33	24.00	52.33	0.33	100
none	0.00	0.00	24.17	0.50	3.83	4.83	3.17	12.17	51.00	0.33	100
Grand Total	99.33	100.00	167.17	48.12	161.80	89.08	15.56	45.61	171.60	1.74	900

Fig. 5. The percentage of responses for the natural initial-consonant visemes.

Stimuli.C	B/P/M	F	D/T/N/L	Z/C/S	ZH/CH/SH/R	J/Q/X	G/K	H	none	blank/error	Total
B/P/M	97.90	1.10	0.10	0.30	0.10	0.30	0.10	0.00	0.00	0.10	100
F	0.50	99.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	100
D/T/N/L	0.17	0.94	59.00	4.67	12.22	8.94	1.67	3.94	8.00	0.44	100
Z/C/S	0.33	0.67	28.78	8.44	27.56	14.67	3.67	1.00	14.67	0.22	100
ZH/CH/SH/R	0.09	1.00	16.64	7.36	48.73	16.64	1.00	0.73	7.73	0.09	100
J/Q/X	0.67	1.17	29.50	10.33	25.83	18.83	1.17	1.17	11.00	0.33	100
G/K	0.17	1.00	55.00	2.00	16.33	5.50	4.83	7.67	7.17	0.33	100
H	0.67	3.33	48.00	3.67	17.00	9.00	7.33	6.67	4.33	0.00	100
none	9.50	10.17	32.17	9.83	11.33	10.00	0.83	1.67	14.33	0.17	100
Grand Total	109.99	118.38	269.18	46.61	159.61	83.88	20.60	22.84	67.23	1.69	900

Fig. 6. The percentage of responses for the synthetic initial-consonant visemes.

(87.72% – 86.11% = 1.61%) 1.61%. For Mandarin “e”, the previous natural–synthetic gap of 23.75% is now reduced to 11.00%. For “i”, the previous gap of 38.64% is now reduced to 19.55%. For “u” and “v” (rounded vowels),

the previous gap of 38.00% is now reduced to 23.20%. Mandarin “er” is the only vowel-ending of which the natural–synthetic gap is not reduced (in fact, the gap is now bigger). However, this may be misleading, because there

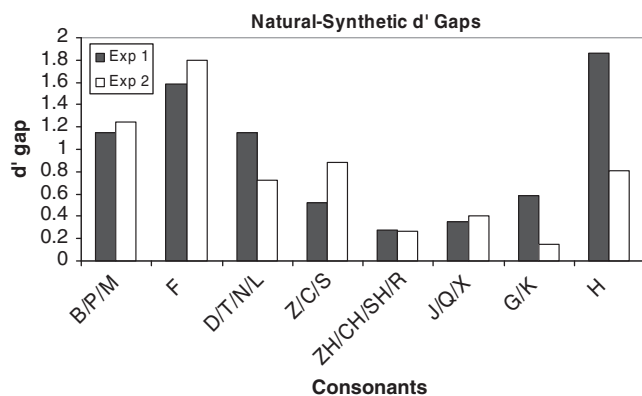


Fig. 7. The natural-minus-synthetic gaps of d' values for both experiments.

was no previous gap: this was a special case where the previous experiment produced 0% accuracies from both natural and synthetic stimuli.

Analogous to analyses with consonants, d' values were calculated for the single-vowel endings. Overall, the previous experiment yielded 2.92 for its natural vowels and 1.59 for its synthetic vowels (a natural–synthetic gap of 1.33). The current experiment yielded 2.82 for its natural vowels and 1.93 for its synthetic vowels (a natural–synthetic gap of 0.89). Synthetic vowels from the current experiment have larger d' values than synthetic vowels from the previous experiment, and the natural-minus-synthetic gap for d' was reduced by a factor of 32.87%. Fig. 10 shows the natural–synthetic gaps of the d' values for both experiments. The natural–synthetic d' -gaps in the current experiment were smaller in 4 of the 5 vowels (compared to the previous experiment). The natural-speech advantage over the synthetic speech was reduced for every vowel except “er”, and this is the same pattern reflected by the results on the percent-correct responses.

Like percent-correct responses, “er” was the only exception in which there was no gap to begin with; in fact, the previous experiment’s d' -gap for “er” was actually negative,

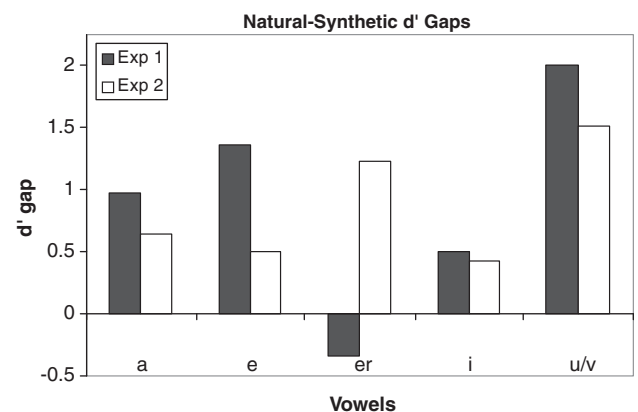


Fig. 10. The natural–synthetic gaps of the d' values for vowels in both experiments.

because the d' for synthetic “er” was actually larger than the d' for natural “er”. The stimulus “er” was only one stimulus out of the 71 unique stimuli, so it only represents a small portion of the stimuli tested. Its results are based on a relatively small number of observations, so each of the data points could potentially influence the results in relatively large proportions.

Another comparison was the correlation between the response pattern yielded by the natural vowels and that yielded by the synthetic vowels. In terms of this correlation, the current experiment’s correlation [$r(38) = 0.94, p < .001$] was higher than that of the previous experiment [$r(38) = 0.82, p < .001$]. Putting the three measures (% correct, d' , and r) together, the results suggested that, comparing synthetic vowels in the current experiment to those in the previous experiment, the current set of synthetic vowels produced responses that were more accurate, and its response patterns are more similar to those of the natural vowels.

Generally, the improvements on the vowels were more impressive than the improvements on the consonants. But taken together, it was still a solid improvement overall.

Stimuli.V	a (an, ang)	e	er	i (in, ing)	u.v (ong, vn)	none (0)	en (eng)	Dip(n).Trip	error/blank	Total
a (180)	30.67	0.00	0.11	0.00	0.17	0.06	0.28	68.50	0.22	100
e (160)	0.69	27.94	0.38	25.25	0.31	7.94	17.81	19.63	0.06	100
er (10)	6.00	2.00	32.00	0.00	0.00	0.00	1.00	59.00	0.00	100
i (110)	0.09	3.82	0.00	65.64	0.18	10.36	0.45	19.27	0.18	100
u.v (250)	0.00	0.04	0.00	0.00	93.68	0.00	0.00	6.12	0.16	100
Grand Total	37.45	33.80	32.49	90.89	94.34	18.36	19.54	172.52	0.63	500

Fig. 8. The average percentage of responses for the natural single-vowel endings.

Stimuli.V	a (an, ang)	e	er	i (in, ing)	u.v (ong, vn)	none (0)	en (eng)	Dip(n).Trip	error/blank	Total
a (180)	45.44	0.22	0.06	0.50	0.22	0.50	0.78	52.06	0.22	100
e (160)	8.94	16.94	1.44	19.19	1.50	7.50	10.19	34.13	0.19	100
er (10)	32.00	12.00	1.00	7.00	2.00	0.00	3.00	43.00	0.00	100
i (110)	3.82	4.00	0.73	46.09	1.45	13.18	3.55	26.27	0.91	100
u.v (250)	1.28	0.20	0.00	1.16	70.48	0.44	0.20	25.88	0.36	100
Grand Total	91.48	33.36	3.22	73.94	75.66	21.62	17.71	181.33	1.68	500

Fig. 9. The average percentage of responses for the synthetic single-vowel endings.

With the exceptions of consonant “z/c/s” and vowel “er”, all of the natural–synthetic gaps in the current experiment are either similar to or smaller than those corresponding gaps in the previous experiment. It was not exactly clear why “z/c/s” is an exception. Although synthetic “z/c/s” (and “er”) from experiment 2 appeared to look more like their natural counterparts compared to those from experiment 1, those specific changes did not help improve performance in the current study. Maybe showing the teeth in Mandarin z/c/s (trying to mimic their fricative/affricate manner) was not as important as their mouth shape in this situation.

3.2.3. Consonant–vowel syllable visemes

Separate analyses showed that Bao’s synthetic consonants and vowels both improved in the current experiment. An additional way to analyze the results was to examine the “Consonant–Vowel visemes” (CV-vis) together in the same syllables. Data can be arranged according to both the consonant-viseme and vowel-viseme at the same time. In this manner, an answer is correct only if the answer matches the whole mouth shape (initial consonant and

vowel ending) or “syllable-viseme”. For example, for the stimulus/word “bi” in this analysis, responses like “pi” and “mi” are counted as correct, but not “ba”, “mu”, “pu”, “ji”, “ti”, etc., because either the consonant-viseme or vowel-viseme is incorrect in these cases. There was also improvement even if the results are analyzed in terms of the viseme of the whole syllable. In this type of analysis, the previous experiment produced a natural–synthetic gap of (33.94% – 15.88%) 18.06% for the percentage of responses, while the current experiment produced a gap of (32.30% – 24.11%) 8.18% (a 54.71% gap-reduction).

As for the d' values, the previous experiment produced a natural–synthetic gap of (1.91–1.17) 0.74, while the current experiment produced a gap of (1.92–1.38) 0.54 (a 26.91% gap-reductions). Correlations were also carried out between the response patterns of the natural and synthetic syllable-visemes for both experiments. The correlation for the current experiment [$r(955) = 0.87, p < .001$] was higher than the correlation for the previous experiment [$r(955) = 0.70, p < .001$]. Although consonants and vowels improved, it is reassuring to know that performance in terms of syllable-visemes (CV-vis within a given word) also

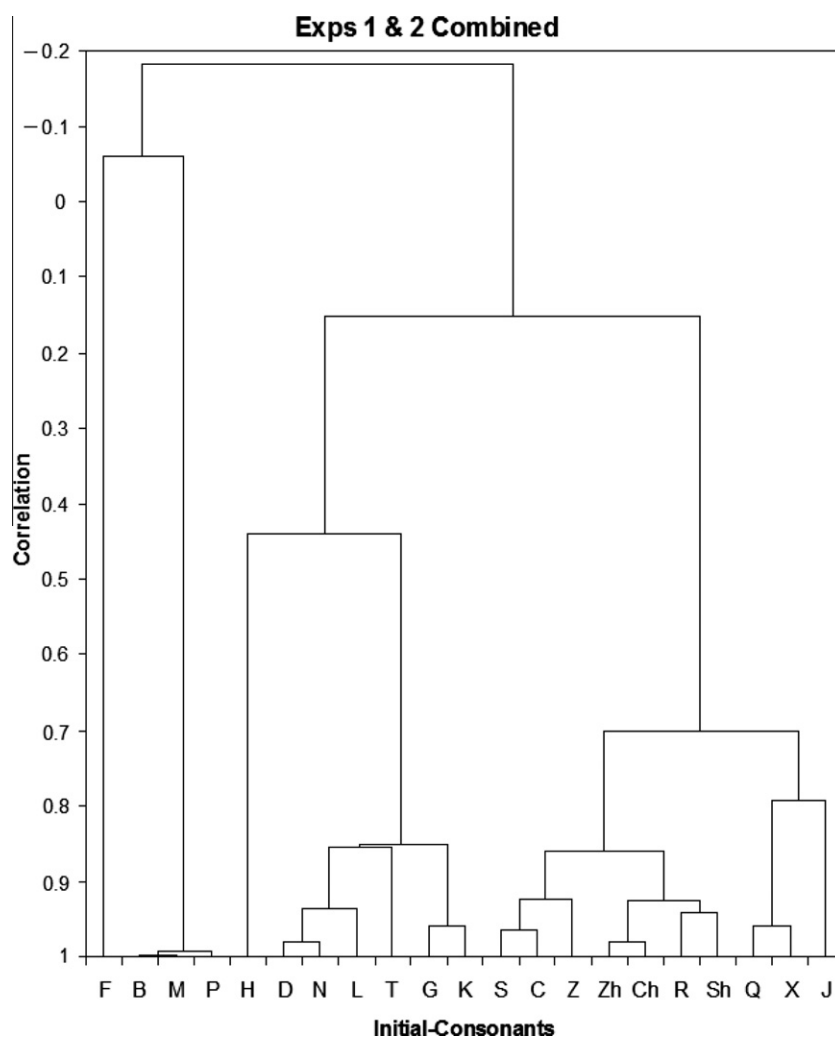


Fig. 11. The clustering analysis for natural consonants (pooled data from both experiments).

improved, because this means that Bao's spoken whole syllables/words also improved. Participants correctly recognized more whole syllables/words in the current experiment than in the previous experiment.

3.2.4. Cluster analysis: Experiments 1 and 2

Clustering analyses were carried out (for initial consonants and single-vowel endings) using correlation coefficient as the metric, and they support the current grouping-arrangement of visemes. Fig. 11 shows the results for the clustering analysis of natural consonants, and Fig. 12 shows the results for the analysis of natural vowels. The data for the natural speech were pooled (combined) across the previous experiment and this current experiment (similar basic patterns of results arose for the natural speech even if the experiments were analyzed separately). Correlation coefficients (based on the pattern of the responses on all the possible phonemes) were calculated between all possible pairings of phonemes, and they established a basis for the degree of closeness between phonemes. Items are in a cluster if they are closer to each other than to other items outside of the cluster. Items and/or clusters were successively grouped together based on their average correlation coefficients, until all items are included in the clustering tree. If $r = 1$, then there is a perfect relationship. If $r = 0$, then there is no relationship. If r is a negative number, then this hints that items/clusters involved may be in different "categories".

An interesting comparison is between the dendrograms of natural stimuli and that of the synthetic stimuli. Fig. 13 shows the results of the clustering analysis for the synthetic consonants in experiment 1, and Fig. 14 shows the results of the analysis for synthetic vowels in experiment 1. Fig. 15 shows the results of the analysis for the synthetic consonants in experiment 2, and Fig. 16 shows the analysis for synthetic vowels in experiment 2. The improvements of the synthetic speech were observed in quantitative measures (percent correct, d-prime, correlation, etc.) and in qualitative comparisons (as shown in comparisons of clustering analyses). Comparing to the clustering analyses of synthetic consonants and vowels from experiment 1, the clustering analyses of synthetic speech from experiment 2 were more similar to those of natural speech.

3.3. Discussion

The natural viseme-categories are consistent with the clustering analyses, Mandarin phonetics, and other previous research. All items in a group have high-positive and significant correlations between each other, and they are the closest to each other. If anything, the current categories were conservative and on the "safe" side. Although it was possible to have fewer visemes, there seemed to be no inherent absolute "cut-off" points. On one end of the extreme, using individual phonemes would dilute a Bao improvement. On the other end of extreme, having too

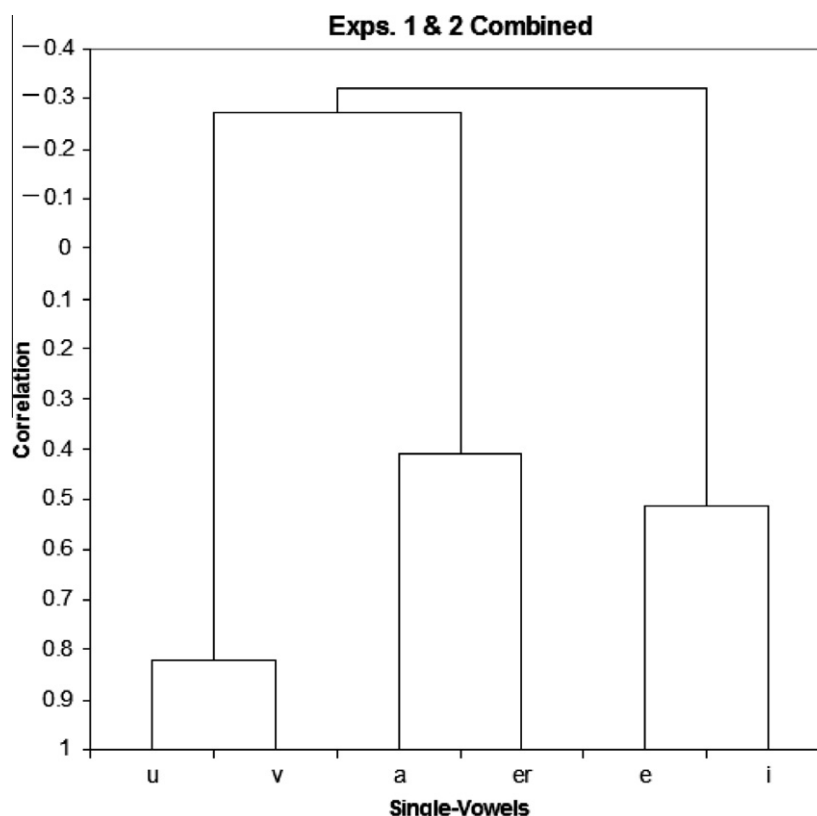
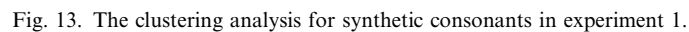


Fig. 12. The clustering analysis for natural vowels (combined/pooled data from both experiments).



For g/k: They were the closest to each other in both the current study as well as in the study by Chen (1991), but g/k in the study by Chen (1991) did not produce high-enough correct responses to be considered as a viseme by themselves. For z/c/s, zh/ch/sh/r, and j/q/x: They were the closest to each other in both the current study as well as in the study by Chen (1991). In the study by Chen (1991), these phonemes all produced relatively low percentages of correct responses, and they were combined (along with g/k

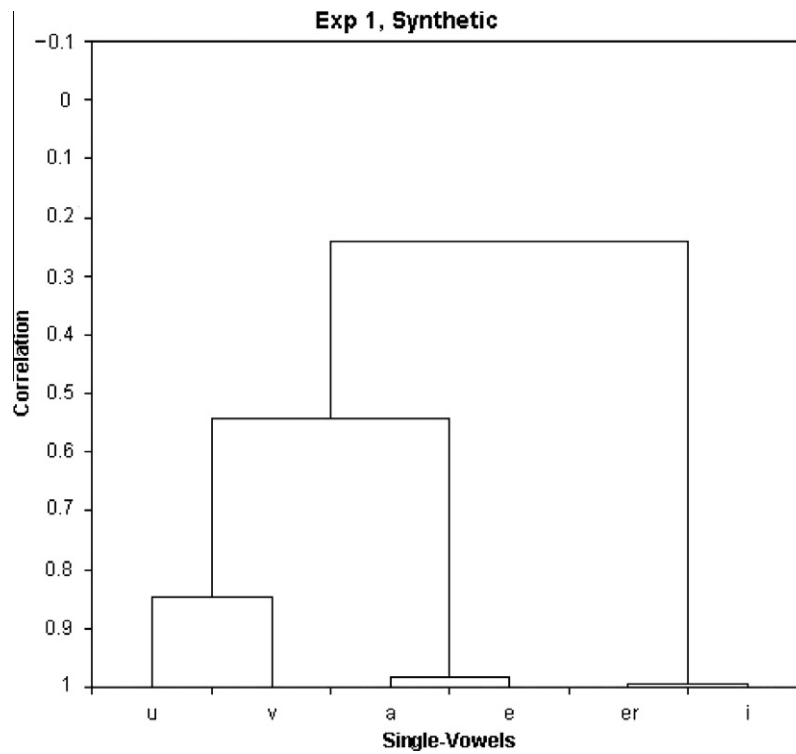


Fig. 14. The clustering analysis for synthetic vowels in experiment 1.

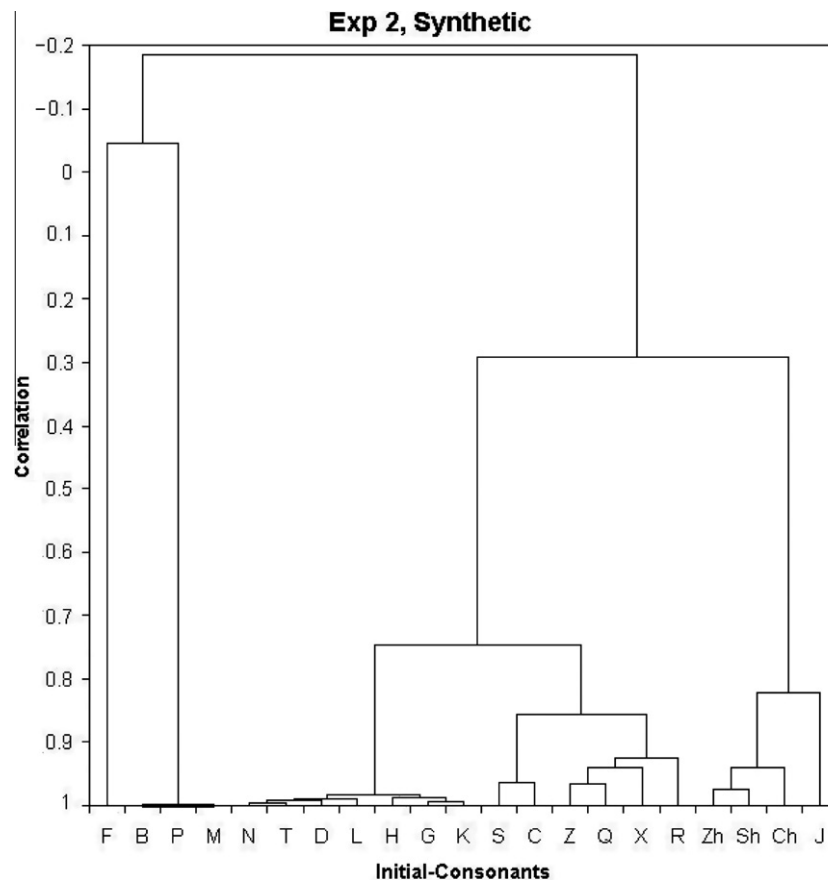


Fig. 15. The clustering analysis for synthetic consonants in experiment 2.

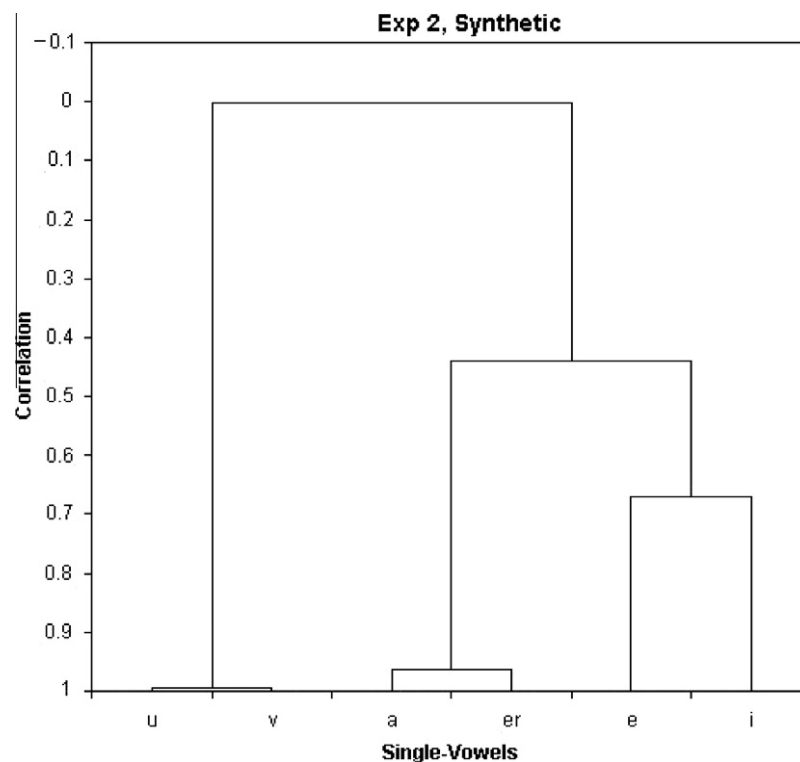


Fig. 16. The clustering analysis for synthetic vowels in experiment 2.

and d/t/n) together to achieve 91.91% correct responses as a viseme. In the current study, they were divided in terms of the closeness of their response patterns, and z/c/s, zh/ch/sh/r, and j/q/x were closer to each other in their respective groups than they were to others (although these three groups were indeed the closest to each other as groups).

Given the fuzzy nature of visual speech (Massaro, 1998, p. 395) and differences across experimental tasks (speaker, stimuli, etc.), perhaps viseme-categories may theoretically span from the level of individual phonemes (an unlikely extreme end for most non-expert speech-readers, who visually perceive differences between visemes better than they perceive differences within visemes) to a level with very few visemes (most perceivers are at least somewhat sensitive to the visual distinction between labials and non-labials). The grouping-arrangement in the current study seemed close to the “minimum unit” of visemes. It is slightly “above” the level of phonemes, but each group only included phonemes that clearly belonged together, both theoretically and based on their response patterns. From the eight categories, it was possible to reduce the number of visemes, but the current study focused on improving the “minimum units” of synthetic visual speech. Given that most of them improved in the current study, this increases the confidence that improvement of the synthetic speech was wide-spread. On the other hand, if the grouping is set at “higher” levels of the continuum (with fewer visemes), it is possible that an improvement might not theoretically be as wide-spread. Having more viseme categories was actually a “harder” test for improving synthetic speech.

It is the same basic story for the natural vowel-visemes. The vowel-visemes from the study by Chen (1992) and those from the current study are not inconsistent with each other. In the current study, only “u” and “v” were grouped together; the rest of the single-vowels were left by themselves, but “e” and “i” were the closest to each other, and “er” and “a” were the closest to each other. In a consistent fashion, the Chen study (1992) put “e” and “i” within the same viseme category and “er” and “a” within another category (which included other vowels and codas). Vowels u and v were not put in the same category by Chen (1992): v by itself had produced 81% correct responses (higher than 75%) and was put in a viseme by itself; u did not produce a percent-correct greater than 75%, and it was not included in any viseme category. But interestingly, Chen (1992) found again that different criteria resulted in different categories, and one of the different types of clustering analysis resulted in u/v being grouped within a viseme (along with “o”, which does not appear as a single-vowel in Mandarin).

Table 3 summarizes Bao’s improvements. Several trends are clear. Both percentages of correct and d’ values yielded by the natural speech are similar or comparable across the two experiments. However, both percent correct and d’ values yielded by the synthetic speech are greater in experiment 2. The natural-minus-synthetic gaps are smaller in experiment 2. The correlations between the response patterns of natural and synthetic speech are higher in experiment 2 compared to experiment 1. In sum, all of the measures suggested that synthetic speech in experiment 2 was more realistic than its previous version in experiment 1. Improvements at the viseme level did not come at the

Table 3

A summary of the results in experiment 1 (part I) and experiment 2 (part II).

			Experiment 1	Experiment 2
Absolute	Natural		11.58%	11.38%
	Synthetic		5.92%	6.54%
	Natural–synthetic gap		5.66%	4.84%
	Natural–synthetic correlation		$r = 0.72$	$r = 0.79$
C-viseme	Percentage correct	Natural	47.71%	50.14%
		Synthetic	32.57%	42.44%
		Gap	15.14%	7.70%
	d' values	Natural	1.89	2.07
		Synthetic	1.05	1.37
		Gap	0.84	0.70
	Natural–synthetic correlation		$r = 0.81$	$r = 0.97$
V-viseme	Percentage correct	Natural	62.04%	57.68%
		Synthetic	35.46%	47.31%
		Gap	26.58%	10.37%
	d' Values	Natural	2.92	2.82
		Synthetic	1.59	1.93
		Gap	1.33	0.89
	Natural–synthetic correlation		$r = 0.82$	$r = 0.94$
CV-viseme or syllable viseme	Percentage correct	Natural	33.94%	32.30%
		Synthetic	15.88%	24.11%
		Gap	18.06%	8.18%
	d' Values	Natural	1.91	1.92
		Synthetic	1.17	1.38
		Gap	0.74	0.54
	Natural–synthetic correlation		$r = 0.70$	$r = 0.87$

expense of hurting the performance at the level of individual phonemes: phoneme-based performance changed little across the experiments.

The current study used 71 unique stimulus-syllables that cover all Mandarin initial consonants, single-vowel endings, and their combinations (they constitute about 17–18% of the entire Mandarin syllabic repertoire). Possible future research can evaluate and improve Mandarin synthetic diphthongs, triphthongs, and/or codas. The work can also be extended to phrases or sentences. The current study achieved solid improvements of Bao's synthetic visual speech, and this was one step towards a realistic Mandarin synthetic talking head.

Acknowledgments

The research and writing of this article were supported by the Federico and Rena Perlino Scholarship Award, the Eugene Cota-Robles Fellowship, and the Psychology Department at the University of California, Santa Cruz (the Doctoral Student Sabbatical Fellowship and the Mini-Grant Research Fellowship). The authors thank Michael M. Cohen for offering expert technical assistance.

Appendix A. The 71 unique word-syllables used in the current experiments (pinyin and tones)

re 4, da 4, zha 1, sa 1, du 2, mu 4, qu (qv) 1, ge 1, zhu 1, ka 3, te 4, er 4, bi 1, shu 1, xi 1, luu (lv) 2, sha 1, gu 3, ga 4, ku 1, e 2, yu (yv) 3, fa 1, nuu (nv) 3, ke 1, ma 1, se 4, he 1, zhe 1, fu

1, za 2, ba 1, ze 2, ti 1, nu 2, ce 4, zu 1, ne 1, ha 1, ca 1, wu 3, hu 1, she 2, ji 1, che 1, mi 2, me 0/5, lu 4, le 4, di 1, ta 1, qi 1, pu 1, de 2, pi 1, bu 4, ju (jv) 1, su 4, cha 1, yi 1, tu 3, pa 1, cu 1, na 2, ru 2, a 1, ni 2, la 1, li 4, chu 1, xu (xv) 1.

References

- Andersson, U., Lyxell, B., Ronnberg, J., Spens, K.-E., 2001. Cognitive correlates of visual speech understanding in hearing-impaired individuals. *J. Deaf Stud. Deaf Educ.* 6, 103–115.
- Bailly, G., RevAret, L., Borel, P., Badin, P., 2000. Hearing by eyes thanks to the “labiophone”: exchanging speech movements. In: COST254 Workshop: Friendly Exchanging Through the Net, Bordeaux, France.
- Bosseler, A., Massaro, D.W., 2003. Development and evaluation of a computer-animated tutor for vocabulary and language learning for children with autism. *J. Autism Dev. Disord.* 33 (6), 653–672.
- Campbell, C.S., Massaro, D.W., 1997. Perception of visible speech: influence of spatial quantization. *Perception* 26, 627–644.
- Caplier, A., Stillitano, S., Aran, O., Akarun, L., Bailly, G., Beauteemps, D., Aboutabit, N., Burger, T., 2007. Image and video for hearing impaired people. *EURASIP J. Image Video Process.* Article ID: 45641.
- Chen, H.-C., 1991. 聽力正常大學生讀話所觀察到的國音聲母視素 (Mandarin consonant visemes speechread by normal hearing college students). 國立臺南師範初等教育學系 (Tainan Teachers College). 「初等教育學報」第四期 (民國八十年), 149–170.
- Chen, H.-C., 1992. 聽力正常大學生讀話所觀察到的國音韻母視素 (Mandarin vowel and diphthong visemes speechread by normal hearing college students). 國立臺南師範學院初等教育學系. 「初等教育學報」第五期 (民國八十一年), 179–202.
- Chen, T.H., Massaro, D.W., 2008. Seeing pitch: visual information for lexical tones of Mandarin-Chinese. *J. Acoust. Soc. Am.* 123 (4), 2356–2366.

- Chen, F., Spinko, V., Shi, D., 2005. Real-time lip synchronization using wavelet network. In: Proc. 2005 Internat. Conf. on Cyberworlds.
- Cohen, M.M., Massaro, D.W., 1990. Synthesis of visible speech. *Behav. Res. Methods Instrum. Comput.* 22 (2), 260–263.
- Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J., de Villiers, J., Tarachow, A., Massaro, D.W., Cohen, M.M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., Soland, C., 1998. Intelligent animated agents for interactive language training. In: STiLL: ESCA Workshop on Speech Technology in Language Learning, Stockholm, Sweden, pp. 163–166.
- Cosi, P., Cohen, M.M., Massaro, D.W., 2002. Baldini: Baldi speaks Italian. In: ICSLP 2002, 7th Internat. Conf. on Spoken Language Processing, September 16–20, Denver, Colorado.
- Fisher, C.G., 1968. Confusions among visually perceived consonants. *J. Speech Hear. Res.* 11 (4), 796–804.
- Gailey, L., 1987. Psychological parameters of lip-reading skill. In: Dodd, B., Campbell, R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*. Lawrence Erlbaum Associates Ltd, pp. 115–141.
- Jackson, P.L., 1988. The theoretical minimal unit for visual speech perception: visemes and coarticulation. *Volta Rev.* 90 (5), 99–115.
- Ladefoged, P., 2001. *A Course in Phonetics*, fourth ed. Heinle & Heinle, Thomson Learning, Inc.
- Lee, W.-S., Zee, E., 2003. Standard Chinese (Beijing). *J. Int. Phon. Assoc. Illus. IPA* 33, 109–112.
- Macmillan, N.A., Creelman, C.D., 2005. *Detection Theory: A User's Guide*, second ed. Lawrence Erlbaum Associates, Inc.
- Massaro, D.W., 1998. *Perceiving Talking Faces: From Speech Perception to A Behavioral Principle*. MIT Press, Cambridge, Massachusetts.
- Massaro, D.W., 1989a. *Experimental Psychology: An Information Processing Approach*. Harcourt Brace Jovanovich, Inc, New York.
- Massaro, D.W., 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ.
- Massaro, D.W., Bigler, S., Chen, T.H., Perlman, M., Ouni, S., 2008. Pronunciation training: the role of eye and ear. In: Proc. Interspeech 2008, Presented at Interspeech 2008, Brisbane, Australia, September 22–26, pp. 2623–2626.
- Massaro, D.W., Light, J., 2004a. Using visible speech for training perception and production of speech for hard of hearing individuals. *J. Speech Lang. Hear. Res.* 47 (2), 304–320.
- Massaro, D.W., Light, J., 2004b. Improving the vocabulary of children with hearing loss. *Volta Rev.* 104 (3), 141–174.
- Massaro, D.W., Light, J., 2003. Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In: Proc. Eurospeech (Interspeech), 8th European Conf. on Speech Communication and Technology, Geneva, Switzerland.
- Massaro, D.W., Liu, Y., Chen, T.H., Perfetti, C.A., 2006. A multilingual embodied conversational agent for tutoring speech and language learning. In: Proc. 9th Internat. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP, September, Pittsburgh, PA), Universität Bonn, Bonn, Germany, pp. 825–828.
- Massaro, D.W., Ouni, S., Cohen, M.M., Clark, R., 2005. A multilingual embodied conversational agent. In: Sprague, R.H. (Ed.), Proc. 38th Annu. Hawaii Internat. Conf. on System Sciences (HICCS'R05). IEEE Computer Society Press, Los Alamos, CA.
- Ming, O., Lin, I.-C., Lee, D.S.D., 1999. Web-enabled speech driven facial animation. ICAT'99. The Virtual Reality Society of Japan.
- Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., Coleman, M., 2006. Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clin. Linguist. Phon.* 20, 621–630.
- Ouni, S., Cohen, M.M., Massaro, D.W., 2005. Training Baldi to be multilingual: a case study for an Arabic Badr. *Speech Commun.* 45 (2), 115–137.
- Ouni, S., Massaro, D.W., Cohen, M.M., Young, K., Jesse, A., 2003. Internationalization of a talking head. In: Proc. 15th Internat. Cong. of Phonetic Sciences (ICPhS'03), Barcelona, Spain.
- Pei, Y., Zha, H., 2006. Vision based speech animation transferring with underlying anatomical structure. In: Narayanan, P.J., et al. (Eds.), ACCV 2006, LNCS, vol. 3851, pp. 591–600.
- Pei, Y., Zha, H., 2007. Transferring of speech movements from video to 3d face space. *IEEE Trans. Visual Comput. Graph.* 13 (1), 58–69.
- Pullum, G.K., Ladusaw, W.A., 1996. *Phonetic Symbol Guide*, second ed. The University of Chicago Press, Ltd, London.
- Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., Jones, C. J., 1977. Effects of training on the visual recognition of consonants. *J. Speech Hear. Res.* 20 (1), 130–145.
- Wang, A.-H., Bao, H.-Q., Chen, J.-Y., 2000. Primary research on the viseme system in Standard Chinese. In: Proc. Internat. Symp. on Chinese Spoken Language Processing, ISCSLP, Beijing, China, October 13–15, pp. 215–218.
- Wang, Z.-M., Cai, L.-H., Ai, H.-Z., 2003. Text-to-visual speech in Chinese based on data-driven approach. *J. Softw.* 16 (6), 1054–1063.
- Wu, Z., Zhang, S., Cai, L., Meng, H.M., 2006. Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar. In: Proc. 9th Internat. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), September 17–21, Pittsburgh, PA, USA.
- Zhou, W., Wang, Z., 2007. Speech animation based on Chinese Mandarin triphone model. In: Sixth IEEE/ACIS Internat. Conf. on Computer and Information Science (ICIS 2007), pp. 924–929.