Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 51 (2009) 604–621

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Read my lips: speech distortions in musical lyrics can be overcome (slightly) by facial information

Dominic W. Massaro, Alexandra Jesse [*,1]

Department of Psychology, University of California, Santa Cruz, CA 95064, USA

Received 22 August 2006; received in revised form 26 May 2008; accepted 26 May 2008

Abstract

Understanding the lyrics of many contemporary songs is difficult, and an earlier study [Hidalgo-Barnes, M., Massaro, D.W., 2007. Read my lips: an animated face helps communicate musical lyrics. Psychomusicology 19, 3–12] showed a benefit for lyrics recognition when seeing a computer-animated talking head (Baldi®) mouthing the lyrics along with hearing the singer. However, the contribution of visual information was relatively small compared to what is usually found for speech. In the current experiments, our goal was to determine why the face appears to contribute less when aligned with sung lyrics than when aligned with normal speech presented in noise. The first experiment compared the contribution of the talking head with the originally sung lyrics versus the case when it was aligned with the Festival text-to-speech synthesis (TtS) spoken at the original duration of the song's lyrics. A small and similar influence of the face was found in both conditions. In the three experiments, we compared the presence of the face when the durations of the TtS were equated with the duration of the original musical lyrics to the case when the lyrics were read with typical TtS durations and this speech embedded in noise. The results indicated that the unusual temporally distorted durations of musical lyrics decreases the contribution of the visible speech from the face.

Keywords: Perception of speech; Perception of music; Audiovisual speech perception; Singing

1. Introduction

Speech science advanced with applications of the information-processing approach, which is based on the assumption that there is a sequence of processing stages and corresponding representations in spoken language understanding. Within the framework of the fuzzy logical model of perception (FLMP), we have argued that speech perception is influenced by multiple sources of information (Jesse et al., 2000/2001). These sources of information are evaluated independently in terms of their support for each response candidate. A single sensory cue can influence several perceived attributes. The duration of a vowel provides information about vowel identity (*bit* versus *beat*), information such as lexical stress (the noun and verb pronunciations of the word *object*), and syntactic boundaries in sentences (word lengthening before a syntactic boundary). A single perceived attribute in speech is usually influenced by several sensory cues, as in the popular example of the many cues for the voicing of a medial stop consonant (Lisker, 1986). Cues for voicing of medial stops include the duration of the preceding vowel, the onset frequency of the fundamental, the voice onset time, and the silent closure interval.

In the FLMP, the obtained information is passed forward through the model in a continuous rather than a categorical fashion. Perhaps the most convincing argument for continuous perception is the realization that no single source of information (e.g. feature) is sufficient for robust perception but rather that multiple sources of information

[*] Corresponding author. Tel.: +31 24 3521371; fax: +31 24 3521213.
E-mail addresses: massaro@ucsc.edu, Alexandra.Jesse@mpi.nl (A. Jesse).

[1] Present address: Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500AH Nijmegen, The Netherlands.

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.05.013

are usually available and influence speech perception. Given the perceptual reality of multiple sources of information, consider the case of two sources of information each being perceived categorically. If the two sources indicate the same speech alternative, there is no benefit of having two sources relative to just one. Either source alone would have been sufficient for perception of that alternative. If the two sources indicate different alternatives, there is also no benefit of having two sources relative to just one. In this conflicting case with just categorical information, there is no principled method to choose between the two alternatives. At best, the perceiver could choose the alternative that corresponds to the source of information that has the best history in predicting the alternative. However, this strategy cannot exploit the utility of the quality of the information that is currently available from each source. We know that context or higher-level constraints are influential but they cannot be beneficial if the stimulus or lower-level information is perceived categorically. Sentential context, for example, would either agree or disagree with the categorization of the speech input. If the sentence context agrees with the speech input, it can provide no additional information. If the sentence context disagrees with the categorization of the speech input, however, the perceiver is faced with a conflicting situation in which the context and acoustic input are inconsistent with one another. It is important to note that these logical arguments are not the only reasons that we reject categorical perception (Massaro, 1987, 1998; Massaro and Stork, 1998).

As an aside, it has been unfortunate that categorical perception is still accepted by some of the speech community and can be found as fact in most introductory textbooks in perception, cognition, linguistics, and cognitive science. We believe that one of the main contributions to this lasting influence is that students of speech perception have equated the categorical symbolic goal of speech perception with the processes that led up to that outcome. No one denies the fact that spoken language communication requires categorical decisions. When a mother points to a set of toys and asks her daughter to bring the ball, the daughter must decide between the ball and a nearby doll. There must be no ambiguity in her response, unless she chooses to request a clarification from her mother. On the other hand, there is no reason why the child has only categorical information about the message. In the FLMP, the support from the different sources of information (e.g. context and perceptual input) for each response alternative is evaluated and then combined to an overall degree of support for each response candidate (e.g. the overall support for BALL) in order to make a decision. The model predicts the probability of a response alternative being selected based on the relative overall support for this alternative compared to the overall support for all competitors (i.e., the overall support for BALL divided by the summed overall support for all other competitors (DOLL, PAUL, etc.)).

Importantly, the model considers information from all available modalities. Contrary to most accounts of speech perception, the understanding of language is a multimodal phenomenon. We now know that sources of information emanate from both the audible speech and the visible mouth movements of the speakers, and that these simultaneously influence speech perception (Massaro, 1987, 1998). Whenever visual speech information (i.e., information from the face of a speaker or gestures) is available, the perceiver uses this source of information for speech recognition. Visual speech contributes to the robust recognition of speech by providing redundant and supplementary information. For example, visual speech provides mainly place of articulation information (Miller and Nicely, 1955), a feature that tends to be most vulnerable in the auditory signal to the addition of noise or a hearing impairment (Massaro and Cohen, 1999; Miller and Nicely, 1955; Summerfield, 1987). Overall, the contribution of visual speech information to the recognition of phonemes is comparable to an increase of the auditory signal by 15dB change (Sumby and Pollack, 1954). The child in the above situation can therefore resolve the auditory ambiguity between DOLL and BALL by seeing the mother speak. The characteristic lip movements during production of bilabial stop /b/ distinguish it visually from the alveolar /d/. However, visual information would contribute less to the distinction between /b/, /p/, and /m/ (i.e. between BALL, PAUL, and MALL).

Although most studies have focused on the segmental contribution of visual speech to recognition, visual speech also provides information about prosodic structure. For example, emotional prosody is better understood with the face presented along with the voice (Ellison and Massaro, 1997; de Gelder and Vroomen, 2000; Massaro and Egan, 1996). Also many turn-taking cues that are essential for an effective interaction in a communication are apparent in the visual modality (see Granström and House, 2005, for a review). Also prominence can be detected in visual speech rather well (Bernstein et al., 1989; Dohen et al., 2004; Dohen et al., 2005; Granström et al., 1999; House et al., 2001; Keating et al., 2003; Lansing and McConkie, 1999; Massaro, 2002; Nicholson et al., 2003; Risberg and Lubker, 1978; Swerts and Krahmer, 2004; Swerts and Krahmer, 2005; Thompson, 1934). Visual speech information contributes to the recognition of lexical tones in Cantonese and Thai (Burnham et al., 2001; Mixdorff et al., 2005) and is informative about lexical stress in English and Swedish (Keating et al., 2003; Risberg and Lubker, 1978). Visual speech information has also been shown to provide information about the intonation contour of an utterance and therefore to aid in the discrimination of statements and echoic questions (Bernstein et al., 1989; Fisher, 1969; Hnath-Chisolm and Kishon-Rabin, 1988; House, 2002; Nicholson et al., 2003; Srinivasan and Massaro, 2003).

Acoustically, this discrimination can be accomplished based on the characteristic changes in the fundamental frequency (F0). Visually, changes in F0 are mostly invisible in articulatory gestures, but correlate strongly with eyebrow and head movements (Cavé et al., 1996; Munhall et al., 2004; Yehia et al., 2002). Viewing the head movements of a speaker, for example, can improve the recognition of words in a sentence (Munhall et al., 2004) compared to seeing a speaker with no head movements. The head movement could contribute to word recognition by aiding the segmentation of the speech signal. Further evidence for the role of visual speech in segmentation comes from studies showing that word and phrasal boundaries can be detected visually (Auer et al., 2004; Risberg and Lubker, 1978; but see Granström et al., 1999). Although visual cues to prosody seem to be communicated mostly by the upper rather than the lower part of the face (Swerts and Krahmer, 2005), and are in general more distributed in the face than cues for segmental identification (Lansing and McConkie, 1999), articulatory information in the lower face (i.e. jaw, cheek, and chin movements) is informative about prosody as well. This is not surprising, since prosodic structure has an influence on the articulation of segments (e.g. articulatory strengthening at boundaries). These articulatory consequences of prosodic structure seem indeed to be used in perception. For example, visual articulatory information is sufficient to provide high rates of correct prominence detection (Dohen et al., 2004; Dohen et al., 2005; Lansing and McConkie, 1999).

1.1. Perception of music and singing

In our theoretical framework, music perception and understanding lyrics are also forms of pattern recognition. Thus, we expect the pattern recognition processes to be similar across speech and music domains. According to the FLMP framework, the sources of information would be different in the two domains but the information processing would be similar. The influence of visual information on auditory perception is not restricted to speech stimuli, but can also found for nonspeech events, such as in the perception of music. For example, seeing a cello player's pluck or bow movements influences the recognition of an auditory stimulus as either being plucked or played with the bow (Saldaña and Rosenblum, 1993), even when the participants are instructed to base their responses on the auditory information alone. In other words, the decision of the perceiver about the musical event is influenced by both modalities. Seeing a performer also provides information about phrasing and emotion (Vines et al., 2006; Dahl and Friberg, 2007). Furthermore, audiovisual information in music perception seems to be processed in the same brain region as for speech perception: Watching a piano player without sound activates for musicians the same region (Hasegawa et al., 2004) as lipreading does (Calvert et al., 1997). And as for speech perception, perceivers of music performance are sensitive to audiovisual asynchrony of the music stimuli (Vatakis and Spence, 2006; Vines et al., 2006).

An interesting occurrence in music is the poor recognition we have of many of the lyrics of both popular and less known songs. Just as for slips of the ear in spoken language (Fromkin, 1971), the misidentification of lyrics is influenced by psychoacoustic and phonetic similarity (Smith, 2003) that leads to segmental recognition errors and misphrasing. For example, the phrase "all of the other reindeer" from the song "Rudolph the red-nosed reindeer" tends to be misheard as "Olive the other reindeer". Or the lyric "There's a bad moon on the rise" from the Creedence Clearwater Revival, "Bad Moon Rising", is perceived as "There's a bathroom on the right" (see e.g., http://www.kissthisguy.com/). Since visual information aids the perception of music as well as the perception of speech, a consequential question is whether visual information also aids the recognition of words when sung rather than spoken. Here, the visual signal could also provide information that facilitates the segmentation of words in a continuous stream as well as the recognition of the word's constituent segments. The question arises how much the process of singing generally alters the realization of prosodic structure and segmental information compared to spoken language and whether any of these alternations affect the informativeness of the visual signal.

Similar to spoken language, music is characterized by its prosodic structure (see Palmer and Hutchins, 2006, for an overview of musical prosody). Music and speech show similarities in their use of prosodic information for the expression of emotions (Juslin and Laukka, 2003). Just as for spoken language, music is structured hierarchically by means of rhythm and pitch (Jackendoff and Lerdahl, 2006). Rhythmic structure is determined by the grouping of parts of the music in hierarchically organized units (e.g. motives, phrases) as well as the assignment of beats according to a hierarchical metrical grid (Lerdahl and Jackendoff, 1983; Todd, 1995). Meter is just like in spoken language an alternation of strong and weak segments (e.g. syllables) and has similar acoustical correlates in music as in speech, such as lengthened duration and higher amplitude of segments in metrical strong rather than in metrical weak positions (Clarke, 1985; Fry, 1955; Sloboda, 1983). Metrical structure influences the production and perception of music (Large et al., 1995; Palmer and Krumhansl, 1990; Schmuckler, 1989). There is evidence that even in the absence of cues to metrical structure in the stimulus material, the perception of music can be guided by knowledge about meter (Palmer and Krumhansl, 1990).

Similarly, the perception of music is also influenced by knowledge about the underlying hierarchical structure. As in language (see, e.g. Fougeron and Keating, 1997), the hierarchical structure is reflected in the degree of phrase-final lengthening (Penel and Drake, 2004; Todd, 1995). Knowledge about the hierarchical structure influences the perception of music, in that lengthening of tones in accordance with the musical prosodic structure (e.g. at phrase-final boundaries) is more difficult to detect than non-predicted lengthening (Repp, 1992, 1998). A musical sequence is better recognized when taken from within a phrase than when straddling a boundary (Tan et al., 1981). This effect is more reliably found for musicians than for untrained listeners (Chiappe and Schmuckler, 1997) and could therefore be evidence for the influence of perception through abstract musical knowledge. On the other hand, there is evidence that musical experience seems not to be necessary for the segmentation of music in phrases. Four-month-old infants are already sensitive to musical structure in that they prefer music fragments that contain a pause, a drop in the pitch contour or lengthening at a phrase boundary rather than in the middle of the phrase (Jusczyk and Krumhansl, 1993; Krumhansl and Jusczyk, 1990; Trainor and Adams, 2000). Recent neuropsychological studies indicate that whether or not music segmentation is driven solely by the perceptual information or also by structural knowledge depends on the level of expertise. Music novices seem to note phrase markers simply as a disruption in the continuity of the music rather than as indicators of a speech-like hierarchical structure as it is the case for musicians (Neuhaus et al., 2006). A more detailed discussion of the degree to which the production and perception of meter and rhythmic structure in general are driven by abstract knowledge, such as mental schemes of hierarchical structures (Todd, 1995), or by a psychoacoustic factors (Repp, 1995) is beyond the scope of the present paper.

It seems that just as in language, phrases are functional units in the processing and planning of music (Chiappe and Schmuckler, 1997; Palmer and van de Sande, 1995). Clicks presented with a musical passage are perceived as migrated towards the phrase boundary (Gregory, 1978; Palmer, 1992; Sloboda and Gregory, 1980). In music production, musical segments from different phrases are less likely to influence each other, i.e., less likely to produce pitch production errors, than when from the same phrase (Palmer and van de Sande, 1995). Just as in spoken language, phrase boundaries are marked in music by multiple cues, often similar to the ones used in speech, such as changes in the melody or pitch contour, or temporal changes, or phrase-final lengthening (Clarke and Baker-Short, 1987; Clarke, 1993; Cuddy et al., 1981; Palmer, 1989; Palmer and Krumhansl, 1987; Palmer and van de Sande, 1995). Musical structure information is not constrained to the auditory signal but can also be perceived from watching a performer. Phrasing can be identified from watching a ballet dancer (Krumhansl and Schenck, 1997) as well as from watching a clarinet player (Vines et al., 2006). Furthermore, in both studies, visual and auditory information converge in cueing the same phrasal boundaries. However, boundaries were detected earlier when provided with visual than with auditory information. Visually, phrase boundaries were communicated by the clarinet player's body sway in accordance with the contour of a phrase. Also motion of fingers and lips can convey temporal information that aids in the detection of temporal changes corresponding to boundaries (note that the music piece used in the Vines et al. (2006) study had no rhythmic meter). In addition, watching the clarinet player breathe as well as characteristic movements of the instrument cued phrase boundaries.

The assignment of phrasal stress also provides evidence for the close connection between the prosodic structure of music and language. In music arrangements for singing, the assignment of phrasal stress follows closely linguistic rules for stress assignment in speech (Palmer and Kelly, 1992). This is independent of whether the music was set to existing lyrics or whether music and lyrics were composed together (Palmer and Kelly, 1992). These stress rules also influence the singing performance: both prosodic structure of the lyrics and musical meter influence the durations of sung syllables (Palmer and Kelly, 1992). Furthermore, the stress assignment rules (e.g. whether a language is stress- or syllable-timed) of a composer's native language influence the musical structure of instrumental compositions (Huron and Ollen, 2003; Patel and Daniele, 2003a,b, 2006).

In summary, speech and music are both hierarchically structured, with similar acoustic correlates and cues to rhythm and boundaries. The perception of musical prosody, just like the perception of speech prosody, seems to benefit from information from the visual modality.

In singing, the articulatory gestures for producing a segment and its acoustic characteristics are modulated by the pitch level the segment has to be sung at. Investigations on the effect of singing on the perception of phonemes have mostly focused on the auditory recognition of vowels. A general trend is that the higher the pitch a vowel is sung at, the less intelligible the vowel (Benolken and Swanson, 1990; Gregg and Scherer, 2006; Scotto di Carlo and Germain, 1985). More specifically, vowels tend to be confused with vowels with higher first formants (Benolken and Swanson, 1990). Consistent with this result, vowels with lower first formants (e.g. /i/) are decreased in intelligibility at lower pitches than vowels (e.g. /a/) with higher first formants (Hollien et al., 2000). The general drop in vowel intelligibility with a raise in pitch can be partially compensated by different singing techniques, e.g. a raising of the larynx increases vowel intelligibility of the soprano singer (Smith and Scott, 1980). Furthermore, placing the vowels in a consonantal context improves the recognition of the vowels, probably through the availability of transitional cues (Smith and Scott, 1980).

For singing at pitches higher than occurring in the normal speaking range, professional soprano singers lower their jaw more to amplify the fundamental frequency by moving the first formant closer to it (Sundberg, 1982). The sung vowel gains in overall amplitude, but the first formant position is therefore less informative about the speech segment than in normal speech. In addition, this technique alters the visual vowel information: Jaw opening in soprano singing changes not as a function of vowel identity as in normal speech but primarily as a function of the intended pitch. However, the usage of jaw opening varies widely among singers (Sundberg and Skoog, 1997). Furthermore, jaw opening is used for low vowels (mainly /a/ and /ɑ/) when the fundamental frequency approaches the first formant while for high vowels (mainly /u/ and /i/) jaw opening is only applied at much higher ranges of fundamental frequency (Austin, 2007; Sundberg and Skoog, 1997). At the highest pitches, all vowels are produced with the same (maximal) jaw opening. To the best of our knowledge, the role of seeing the jaw opening on audiovisual perception of vowels has not been investigated.

Changes in vowel intelligibility due to singing are mostly a problem in high-pitch female singing voices, such as sopranos. Male singers have a vowel intelligibility problem for a set of vowels with low formant frequency sung at the high pitches of a tenor or baritone range (Sundberg, 1982). However, male singers also need to alter their articulatory configurations for singing in order to improve their intelligibility. The frequency range of male singers often falls in the same range as that of an orchestra. To avoid masking by the orchestra, the male singers try to gain in amplitude by moving the third, fourth, and fifth formants closer to each other (Sundberg, 1974). To obtain this "singer's formant," the singer typically lowers the larynx (Sundberg, 1974). On the one hand, this affects the first and the second formant (Sundberg, 1982) and might impact the recognition of vowels, on the other hand, this increases the amplitude of the singer and therefore increases his overall intelligibility (Sundberg, 2003). The extent to which a singer uses this technique of a singer's formant varies. The singer's formant is louder when the singer sings alone than when singing in a choir (Rossing et al., 1986).The singer's formant is stronger for trained than for untrained singers. For trained singers, the singer's formant is also stronger when singing than when speaking (Omori et al., 1996); but see Lundy et al. (2000). Generally, professional singers can produce sounds at higher amplitudes than nonsingers, for example, tenors produce 10–15 dB higher amplitudes than nonsingers (Titze and Sundberg, 1992).

The production of consonants is also altered by singing (McCrea and Morris, 2005a,b, 2007). Voiceless plosives are produced with longer voice onset time in singing than in speaking, but no difference was found for voiced plosives (McCrea and Morris, 2005a). However, this result is to be interpreted with caution since a follow-up study showed a longer voice onset time in singing than in speaking for voiced plosives, but found the reversed pattern for voiceless plosives (McCrea and Morris, 2007). Further research needs to clarify the causes of these differences between the experiments. In addition, the influence of pitch on consonantal intelligibility in singing needs to be investigated. In a reading task, the production of voiceless plosive is influenced by the pitch level (McCrea and Morris, 2005b). Voiceless plosives tend to have shorter voice onset time when read at a higher rather than lower fundamental frequency. To the best of our knowledge, a comparable study for singing has yet to be conducted.

In summary, the articulation of speech segments is altered when sung. This is especially likely for singing at high pitch levels or when in accompaniment from a loud orchestra. However, these changes are mostly found for professional singers (Omori et al., 1996; Titze and Sundberg, 1992) and are modulated by style. For example, professional male country singers show no sign of a singer's formant (Cleveland et al., 2001). Furthermore, the singing in country and Broadway style is more similar to speaking than classical singing (Cleveland, 1994; Stone et al., 2003).

1.2. Audiovisual perception of sung lyrics

Music perception is just like speech perception multimodally influenced. Given the similarities between speech and music in terms of their hierarchical prosodic structures as well as in terms of shared cues to rhythm and boundaries, it is not surprising that seeing a musician provides prosodic information for phrasing (Vines et al., 2006). Seeing a singer might therefore also aid in segmenting the continuous input stream of lyrics in phrases and word. Singing can change, however, the articulatory gestures with which a segment is produced when sung rather than when spoken. The degree of to which this modification occurs in singing depends on the pitch level, the professionalism of the singer, environmental variables (e.g. if singing solo or with an orchestra), and singing style. The question is whether given these modifications, visual singing is still informative for perceiving the lyrics of the songs.

Hidalgo-Barnes and Massaro (2007) demonstrated that visual speech information contributes to the recognition of words in phrases when these are sung instead of being spoken. The articulatory movements of a computer-animated talking head (Baldi®; see Massaro, 1998, for a detailed description) were aligned with an audio recording of the sung lyrics. The articulatory movements of the talking head were, however, not specifically modified for singing. It was assumed that for this singing style and this singer the visual movements would not differ much from speaking. However, the timing of the articulatory movements were exactly aligned to the singing voice. Seeing the singer improved the recognition of the words above and beyond simply listening to the singer, but the contribution of visual information was somewhat smaller than usually found for spoken language (Jesse et al., 2000/2001). However, the word recognition in singing had not been directly compared to word recognition of the same lyrics in speaking. Therefore, it cannot be determined whether this particular set of lyrics simply did not favor a visual contribution or whether the act of singing lowered the informativeness of the visual signal.

The present study investigates this issue by allowing a comparison of the influence of the face in understanding sung lyrics to the case where lyrics are spoken. Therefore, the first experiment tests word recognition performance for lyrics when spoken rather than sung. This time, the lyrics were spoken but with the durations of the individual segments kept the same as in the original singing. Since the audiovisual benefit found was still smaller than what is usually found for spoken word recognition, the final three experiments investigated the influence of the temporal distortion that singing had on the articulation of the lyrics, by comparing the recognition of words in the spoken versions of the lyrics with the original durations as in singing to recognition of words normally spoken. Auditory noise was added to the normal speech condition. Although other differences between singing and speech may exist (such as changes in formant structure in the vowels), we concentrated on the influence of durational changes on word intelligibility. In summary, this study follows up why the spoken sentences in noise (Jesse et al., 2000/2001) benefited so much more by the presence of a face than did the sung musical lyrics in Hidalgo-Barnes and Massaro (2007).

2. Experiment 1

The first experiment investigates the role of durational changes in singing on lyrics comprehension. To see whether the audiovisual benefit found in the Hidalgo-Barnes and Massaro (2007) study that is smaller than what is usually found for speech is due to the lyrics materials or rather inherent to the comprehension of words when sung, we tested word recognition for the lyrics materials when spoken. A spoken version of the original lyrics as in the Hidalgo-Barnes and Massaro (2007) study was modified so that the durations of the individual speech segments were altered to match the durations used in the singing version of the lyrics.

2.1. Method

2.1.1. Participants

Twelve undergraduate students from the Psychology participant pool at the University of California at Santa Cruz participated in this experiment to fulfill a course requirement. None of the participants reported any hearing or language deficits. All participants were native speakers of English.

2.1.2. Material

The 34 phrases tested in this experiment were verses taken from the lyrics of the song "The Pressman" by Primus (1993). The duration of the phrases was about 1–3 s. Appendix A provides the lyrics of the song. Sound samples can be found at http://mambo.ucsc.edu/psl/primus.html. This song had been selected by Hidalgo-Barnes and Massaro (2007) as it was thought to be difficult to understand and lyrics were not emotionally antagonizing. A sample phrase is "by the light of lamp I sit to type". Details about the song can be found in Hidalgo-Barnes and Massaro (2007).

In the Hidalgo-Barnes and Massaro (2007) study, the 34 acoustic phrases were synchronized to the articulatory movements of the talking head, Baldi, with a speech alignment program in the Center for Spoken Language Understanding (CSLU) speech toolkit (http://cslu.cse.ogi.edu/toolkit/). The program takes a text file and corresponding digital audio "wav" file and provides a rough approximation of the location of the phonemes in the sound sample. The alignment was then hand adjusted such that Baldi visually portrayed each phoneme that occurred in each word of the lyrics at the same time and for the same duration as the auditory phonemes present in the song recording. Note that Baldi has been extensively used for investigating the perception of visual and audiovisual speech. The quality of his visual speech has been shown to resemble closely those of a human speaker (see, e.g. Massaro, 1998; Ouni et al., 2007). Examples of these test trials can be found at http://mambo.ucsc.edu/psl/primus.html. In the present temporally yoked experiment, instead of using the original singing audio track, the lyrics and the durations of its phonemes as found in the singing were used as input to the Festival text-to-speech synthesis (TtS), which produced the auditory speech rendition of the lyrics. These same phonemes and durations were also used to control Baldi's articulation in the visual and bimodal speech conditions. Baldi visually portrayed each phoneme that occurred in each word of the lyrics with the same visible-speech movements at the same time and for the same duration as the auditory phonemes present in the original song recording. This was done for the visual as well as for the AV modality condition. To summarize, the Hildago-Barnes and Massaro experiment aligned Baldi with the song's original acoustic lyrics; in this experiment, Baldi was aligned with the audio track of Festival TtS spoken at the original duration of the song's lyrics.

2.1.3. Apparatus

The stimuli were presented on PCs running the Windows 2000 operating system with Open-GL video cards, 17-in video monitors, and a sound-blaster audio card. All of the experimental trials were controlled by an application with PSL Tools (http://mambo.ucsc.edu/psl/tools/tutorial.html) in the speech toolkit from the CSLU (http://cslu.cse.ogi.edu/toolkit/). Baldi's image subtended a visual angle of roughly 15 degrees in height and 7.5 degrees in width. Fig. 1 shows an example of a trial. The auditory speech was presented at a comfortable listening intensity and was held constant for all participants.

2.1.4. Procedure

The experiment consisted of two blocks with an overall total of 204 trials. In each of two blocks, the participant was shown each of the 34 samples once in each of the three modalities, auditory-only (A), visual-only (V), and audiovisual (AV), for a total of 102 trials. These 102 unique trials were randomly presented within each block. The two blocks sessions were separated by a 5-min break. The trials were self-paced and each session took about 30 min to complete. Participants were instructed to listen and watch the computer monitor during the sentence presentation on each trial. The task was to type as many words as they thought they had understood. Participants were informed that on some trials the sentence the speaker said was

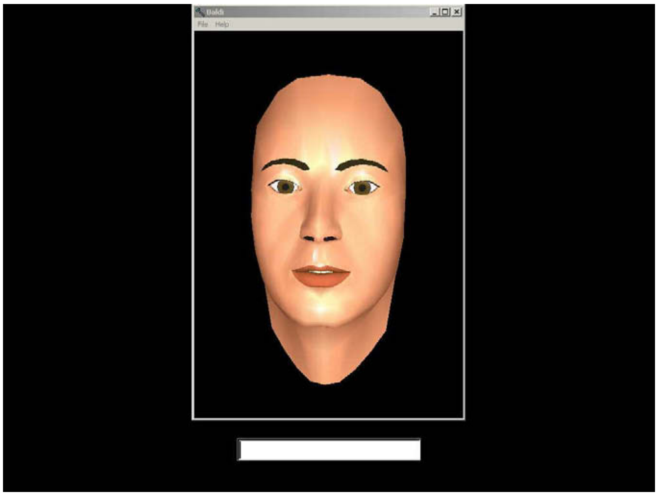

Fig. 1. Screen shot of Baldi, as viewed in the experiment. The white box below Baldi shows the words as they are typed in by the participant. On auditory alone trials, only the box was present on the screen.

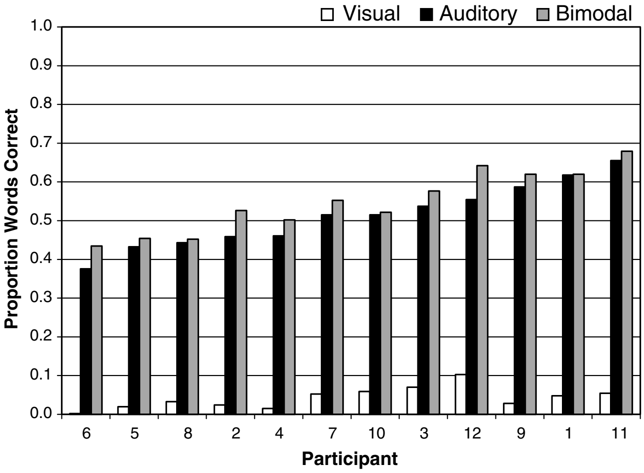

Fig. 2. The individual participant results for the three conditions in Experiment 1 when the durations of the TtS were equated with the duration of the original musical lyrics: A, V, and AV. The results are arranged from left to right according to performance on the A condition.

stretched out. The sentence on each trial had no specific connection to any other trial. Participants were able to see what they typed on the screen and allowed to make any corrections before hitting the enter key to go on to the next trial.

2.2. Results

The proportion of words correctly recognized regardless of position in the verse was computed for each participant under each experimental condition (pooled across verse). For this analysis, all responses had been corrected for obvious spelling errors. An analysis of variance was carried out on proportion of words correctly recognized as dependent variable and block and modality condition (A, V, and AV) as within-subject independent variables.

The results show that the presence of the face did indeed help in the comprehension of the spoken lyrics. The participants were able to understand 51% and 4% of the lyrics with just the auditory and visual lyrics, respectively, whereas performance was 55% in the audiovisual presentation ($F(2, 22) = 585$, $p < 0.001$). A specific planned comparison between the A and AV conditions was statistically significant ($F(1, 11) = 22.9$, $p < 0.001$). Performance also improved from 34% correct in the first session to 39% in the second session, $F(1, 11) = 21.9$, $p < 0.001$. The amount of improvement was somewhat greater for the A condition than for the AV condition, and the V condition showed the least improvement, $F(2, 22) = 5.03$, $p < 0.02$.

Fig. 2 gives the individual participant results for the three conditions. As can be seen in the figure, there is a reasonable range of performance across the 12 participants, and some persons benefited more from the presence of the face than others. However, each participant showed an overall advantage of having Baldi aligned with the verses relative to the single modality conditions.

A second analysis of variance was carried out on the proportion of words correctly recognized as the dependent variable and with modality condition and verse as the independent variables. In addition to an effect by modality, $F(2, 46) = 1109$, $p < 0.001$, there was a large effect of verse, $F(33, 363) = 11.8$, $p < 0.001$, and an interaction between verse and modality conditions, $F(66, 726) = 6.52$, $p < 0.001$. Fig. 3 shows the individual verse results for the three conditions. As can be seen in the figure, there is a fairly broad range of performance across the 34 verses, and the face was more effective in some of the verses than others. Appendix A lists the lyrics for each of the 34 verses.

Finally, an analysis compared performance for the temporally distorted presentation conditions in the present study with the sung presentations in the Hidalgo-Barnes and Massaro (2007) study (see Fig. 4). An ANOVA with presentation condition as a between-subject factor comparing the auditory alone to the AV condition indicated that the benefit of the visual information did not differ across the two presentation conditions, $F(1, 23) = 1.446$, $p = 0.24$. Similarly, there was no difference in the size of the visual benefit across the two experiments, $F(1, 23) = .066$, $p = 0.79$.

2.3. Discussion

The results showed that there did not seem to be anything unique about the musical lyrics materials that were sung, which was responsible for a small contribution of visible speech. When we presented these lyrics by a TtS synthesizer at the same distorted durations, the same small advantage was observed. In the next three experiments, we tested the role of temporal distortions on lyrics through singing by comparing the presence of the face when the durations of the TtS were equated with the duration of the original musical lyrics to the case when the lyrics were read with typical TtS durations and this speech embedded

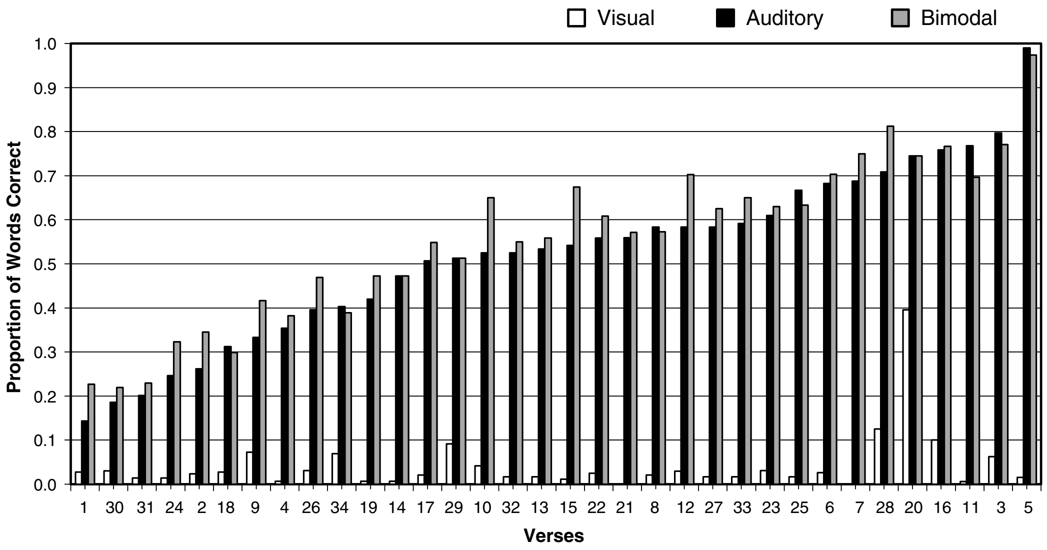

Fig. 3. The individual verse results for the three conditions in Experiment 1: A, V, and AV. The results are arranged from left to right according to performance on the A condition.

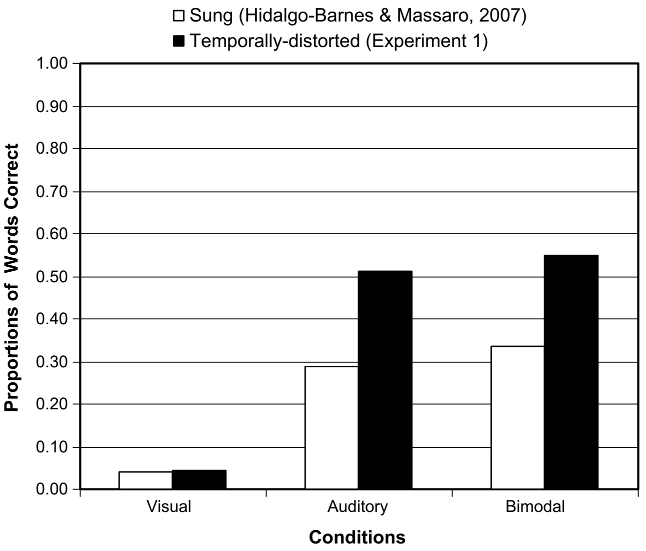

Fig. 4. The average results for the three modality conditions (A, V, and AV) for the sung presentation in the Hidalgo-Barnes and Massaro (2007) experiment and the temporally distorted presentation in Experiment 1.

in noise. Note that spoken rather than sung materials were used throughout these experiments to isolate the influence of temporal distortion as introduced through singing on the intelligibility of lyrics.

3. Experiment 2

3.1. Method

3.1.1. Participants

Twelve undergraduate students from the same population as in the first experiment were tested. None of the students had participated in the previous experiment. All of them reported to be native English speakers with no hearing deficits.

3.1.2. Material

The verse material was the same as in the first experiment. In addition to the temporally distorted presentations, a set of normal speech stimuli for these phrases was created by using Festival TtS. These stimuli were not temporally altered to match the tempo and rhythm of the song, but rather represent normal-duration speech, as if the lyrics of the song were simply read. For V and AV presentation conditions, the articulation of Baldi was driven by the synthetic speech engine. Auditory white noise (–5dB SNR) was added to the A and AV trials.

3.1.3. Procedure

The experiment consisted of a total of 204 trials. Each of the 34 verses was presented under each of the three modality conditions for each of the two presentation conditions (i.e. when the durations of the TtS were equated with the duration of the original musical lyrics and when the lyrics were read with typical TtS durations and this speech embedded in noise). Trials were completely randomized. The experiment was self-paced. Instructions and setup of the individual trials were the same as for the first experiment. The apparatus was also the same as before.

3.2. Results

The scoring and the dependent variable (percentage of correctly identified words for each verse) were the same as in the previous experiment. An analysis of variance with modality (A, V, and AV) and presentation condition (normal, distorted speech) as within-subject independent variables showed a significant effect of modality condition on performance ($F(2, 22) = 1345$, $p < 0.001$). Fig. 5 shows the performance for individual participants in each modality and presentation condition. Only 8 out of the 12 participants showed an audiovisual benefit for the temporally distorted version. Fig. 6 shows the performance by verses.

The main effect of presentation condition was not significant, $F(1, 11) = 1.70$, $p = 0.22$ (46% of all words presented in normal speech in noise and 47% of all words presented distorted were recognized). However, there was a significant interaction effect of modality and presentation condition, $F(2, 22) = 117$, $p < 0.001$. While there was no difference in performance depending on presentation condition for speech presented V, there was a significant difference for the A and AV presented speech. There was a significant advantage of AV versus A (0.86 versus 0.46) in the speech in noise presentation condition, $F(1, 11) = 366$, $p < 0.001$, but only a marginal difference in the temporally distorted condition (0.63 versus 0.66), $F(1, 11) = 2.81$. $p = 0.12$. Fig. 5 shows that this audiovisual benefit for speech in noise was found for all participants. Although contribution of visible speech differed for the two presentation conditions, there was no difference between them for the visual-alone trials (0.075 versus 0.088), $F(1, 11) = 1.32$, $p = 0.27$.

A second analysis of variance with modality condition, presentation condition, and verse was carried out on proportions of words correctly recognized. There was an effect of modality, $F(2, 22) = 1193.87$, $p < 0.001$, and of verse on performance, $F(33, 363)$=8.31 , $p < 0.001$, but not of presentation condition, $F(1, 11) = .725$, $p = 0.58$. All interactions between the three independent variables were also significant ($p < 0.001$).

3.3. Discussion

In Experiment 2, we found a much smaller advantage of the presence of the face when the durations of the TtS were equated with the original musical lyrics than when typical

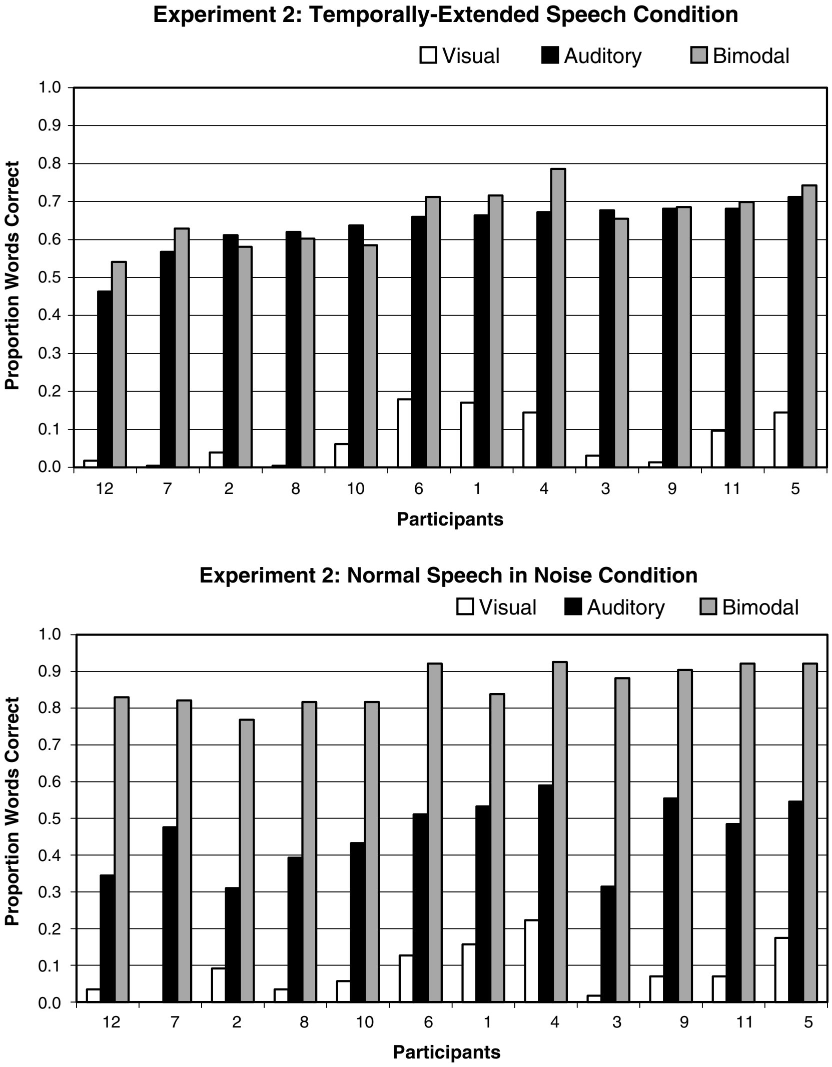

Fig. 5. The individual results for each participant for the three modality conditions for the temporally distorted speech and normal-speech-in noise conditions in Experiment 2: A, V, and AV. The results are arranged from left to right in the temporally distorted speech condition according to performance on the A condition. The normal-speech-in noise condition maintains this ordering.

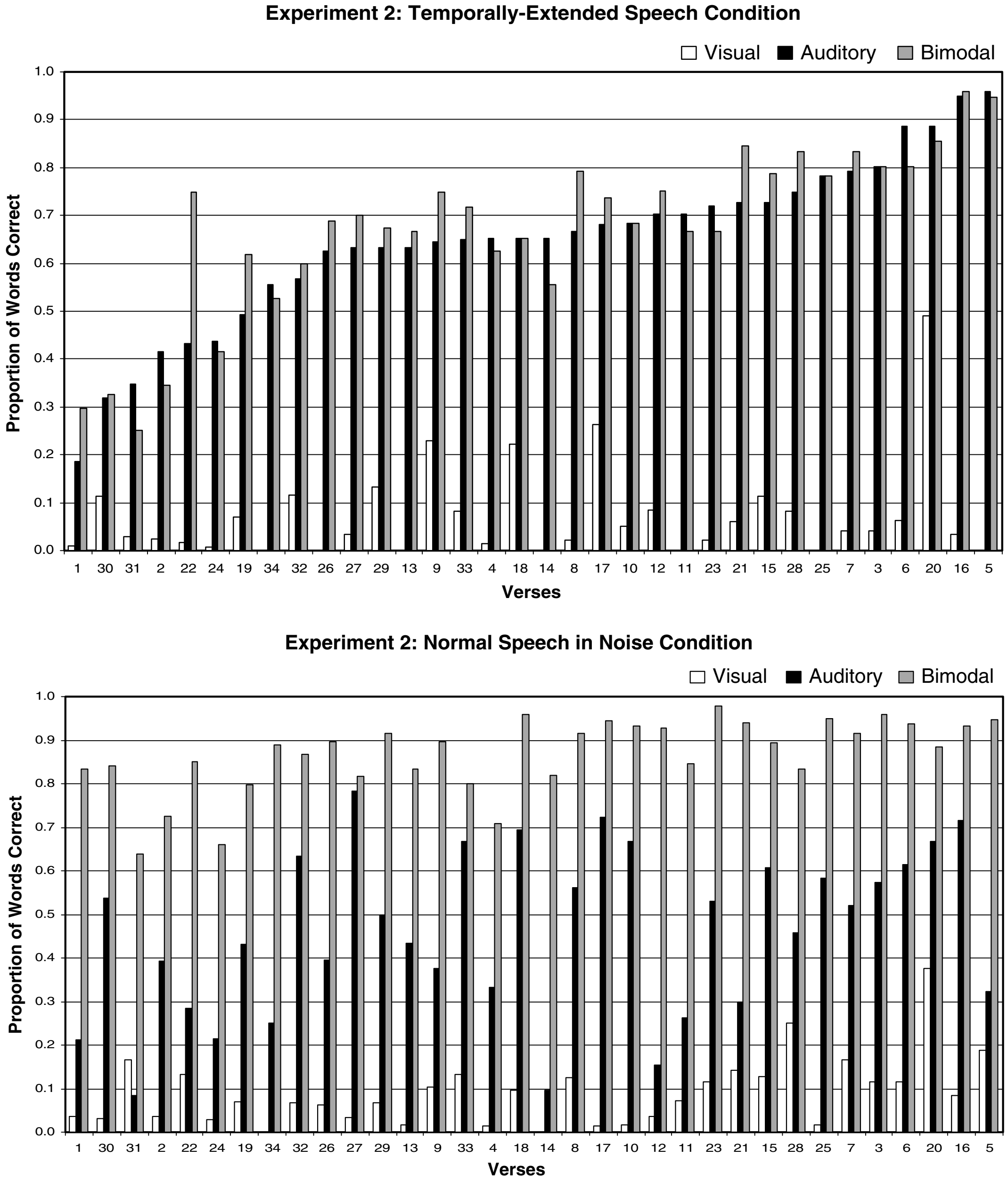

Fig. 6. The individual results for verses for the three modality conditions for the temporally distorted speech and normal-speech-in noise conditions in Experiment 2: A, V, and AV. The results are arranged from left to right in the temporally distorted speech condition according to performance on the A condition. The normal-speech-in noise condition maintains this ordering.

TtS durations were used with the auditory speech embedded in noise. This is despite the fact that the performance on V trials was comparable in both conditions. However, the level of performance on A trials differed between the presentation conditions. Namely, performance was lower for A presentations of speech in noise than of temporally distorted speech. Therefore, the next experiment decreased the amount of noise added in the speech in noise presentation condition to equate the level of performance to that found for temporally distorted speech. (As emphasized by Ouni et al. (2007), measures that use the performance in the AV conditions relative to the A condition (e.g. Sumby and Pollack, 1954) do not necessarily give valid measures of the influence of visible speech.)

4. Experiment 3

4.1. Method

4.1.1. Participants

Fourteen undergraduate students participated in this experiment for course credit. None of them had participated in the previous experiments. Again, all participants indicated to be native English speakers with no hearing deficits.

4.1.2. Material

The stimuli material and procedure were the same as used in the previous experiments. Procedure and

Apparatus were also as before. The signal-to-noise ratio was set to 0dB SNR.

4.2. Results

Again, the scoring and the dependent variable (percentage of correctly identified words for each verse) were the same as in the previous experiments. An analysis of variance on percentage of correctly identified words with modality (A, V, and AV) and presentation condition (normal, distorted speech) as within-subject independent variables was carried out with subjects as a random factor. The analysis revealed a significant effect of modality on performance, ($F(2, 26) = 402$, $p < 0.001$), but only a marginal effect of presentation condition, ($F(1, 13) = 2.79$, $p = 0.12$). However, there was a significant interaction effect between these two variables, ($F(2, 26) = 3.69$, $p < 0.05$). A second analysis of variance with verses as a random factor shows also a significant effect of modality on performance ($F(2, 26) = 401.34$, $p < 0.001$). Unlike for the analysis with subjects as random factor, the verse analysis reveals a significant effect of presentation condition ($F(1, 13) = 5.99$, $p < 0.05$). The interaction between these two variable was also significant ($F(2, 26) = 5.93$, $p < 0.01$).

Specific comparisons show that adding the face during presentation improves the recognition of words for the normal speech presented in noise, namely from 52% for the A to 58% for the audiovisual presentation, $F(1, 13) = 21.9$, $p < 0.001$, but only marginally significant for the temporally distorted speech (55% versus 57%), $F(1, 13) = 4.09$, p = .06. Fig. 7 shows the performance for individual participants in each modality and presentation condition; Fig. 8 shows performance for individual verses for each condition. While only 9 participants showed an audiovisual benefit in the temporally distorted presentation condition, 11 out of the 12 participants showed such an AV benefit for the speech in noise condition. The percentage of correctly identified words for the V condition was 4.7% for the

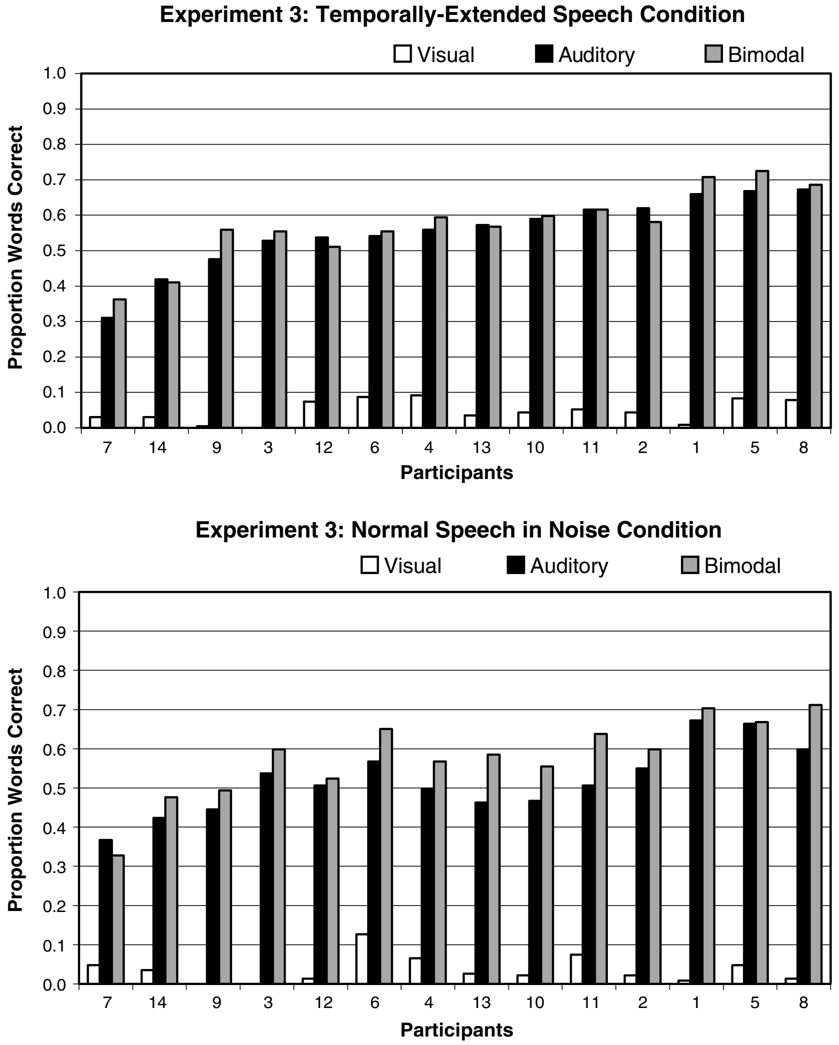

Fig. 7. The individual results for each participant for the three modality conditions for the temporally distorted speech and normal-speech-in noise conditions in Experiment 3: A, V, and AV. The results are arranged from left to right in the temporally distorted speech condition according to performance on the A condition. The normal-speech-in noise condition maintains this ordering.

temporally distorted visual speech and 3.6% for the normal speech, $F(1, 13) = 2.12$, $p = 0.16$.

4.3. Discussion

The results of Experiment 3 replicate Experiment 2 in that an audiovisual benefit is only found for speech presented in noise but not for temporally distorted speech. Again, the performance on V trials was comparable across presentation conditions. Experiment 4 is a direct replication of Experiment 3.

5. Experiment 4

5.1. Method

5.1.1. Participants

Thirteen undergraduate students participated in this experiment for course credit. None of them had participated in the previous experiments. Again, all participants indicated to be native English speakers with no hearing deficits.

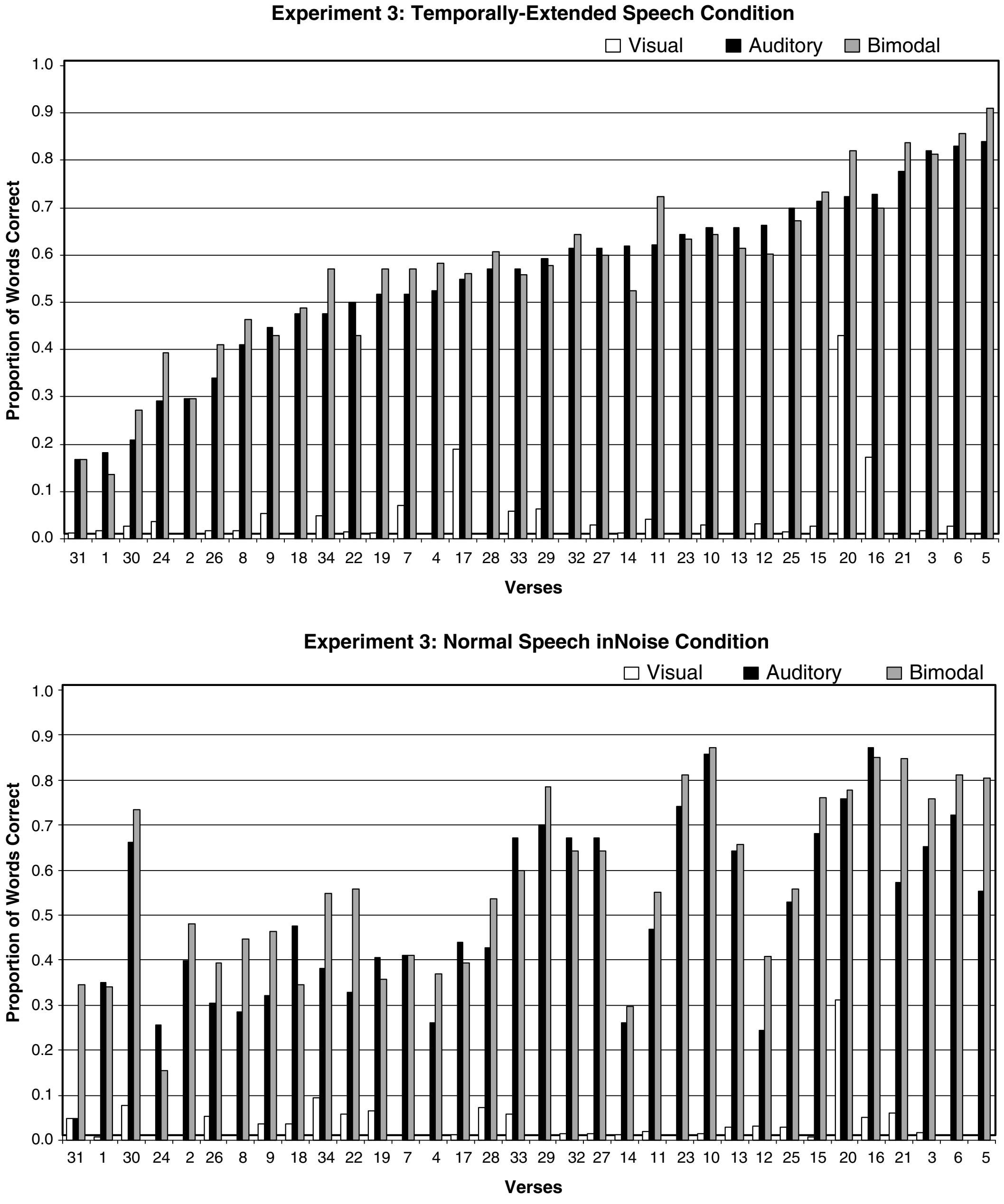

Fig. 8. The individual results for verses for the three modality conditions for the temporally distorted speech and normal-speech-in noise conditions in Experiment 3: A, V, and AV. The results are arranged from left to right in the temporally distorted speech condition according to performance on the A condition. The normal-speech-in noise condition maintains this ordering.

5.1.2. Material

The stimuli material was the same as used in the previous experiments. Procedure and Apparatus were also as before. Auditory white noise was added again to the normal duration speech condition. The signal-to-noise ratio was set to 0dB SNR.

5.2. Results

An analysis of variance on percentage of correctly identified words with modality (A, V, and AV) and presentation condition (normal, distorted speech) as within-subject independent variables was carried out. The scoring and the dependent variable (percentage of correctly identified words for each verse) were the same as in the previous experiments. The analysis revealed a significant effect of modality on performance ($F(2, 24) = 817$, $p < 0.001$), a significant effect of presentation condition ($F(1, 12) = 9.45$, $p < 0.009$), and a significant interaction effect between these two variables ($F(2, 24) = 9.33$, $p < 0.001$). A second analysis of variance on verses as random factor showed also a significant modality condition effect ($F(2, 24) = 809.05$, $p < 0.001$) and of presentation condition ($F(1, 12) = 4.87$, $p < 0.05$), as well as a significant interaction of both factors ($F(2, 24) = 12.04$, $p < 0.001$).

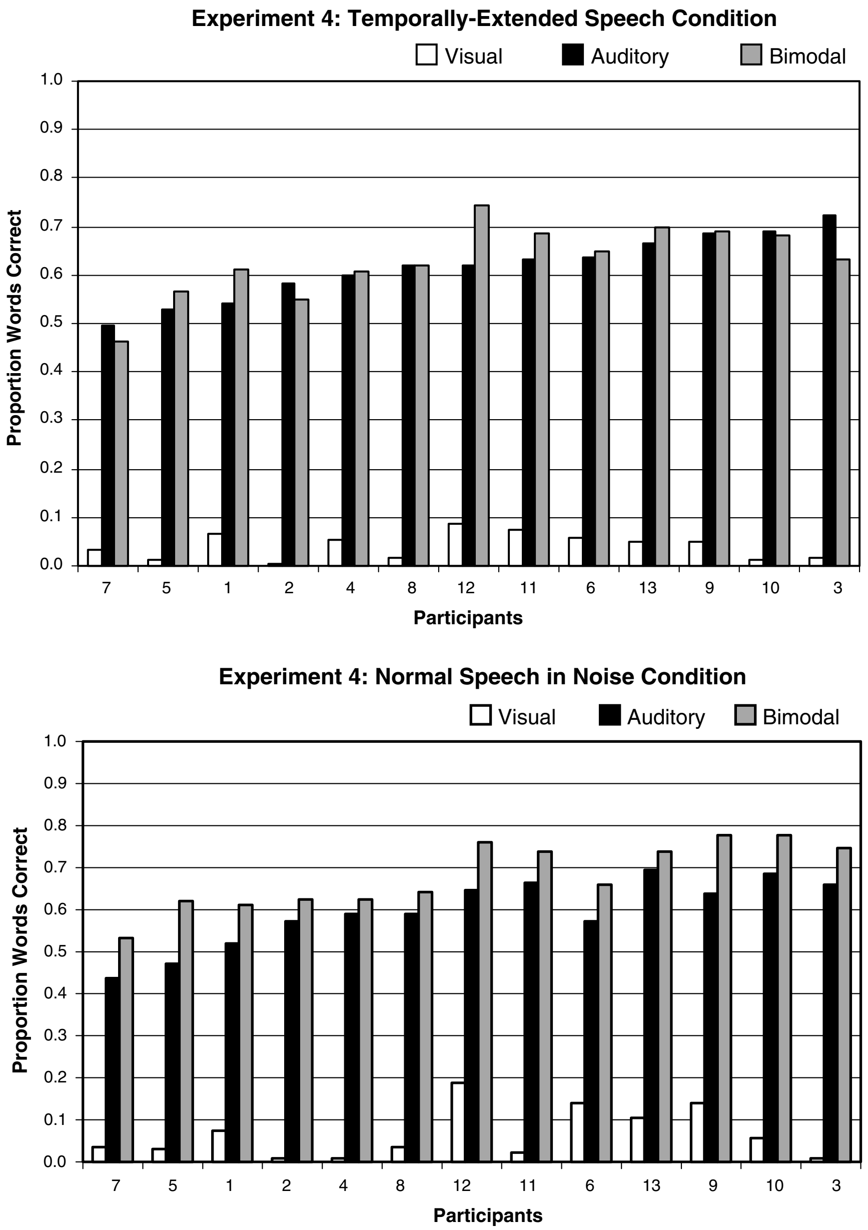

Fig. 9. The individual results for each participant for the three modality conditions for the temporally distorted speech and normal-speech-in noise conditions in Experiment 4: A, V, and AV. The results are arranged from left to right in the temporally distorted speech condition according to performance on the A condition. The normal-speech-in noise condition maintains this ordering.

Specific comparisons show that adding the face during presentation improves the recognition of words for the normal speech presented in noise, namely from 59% for the A to 68% for the audiovisual presentation, $F(1, 12) = 78.9$, $p < 0.001$, but not for the temporally distorted speech (62% versus 63%), $F(1, 12) = 0.88$, $p = 0.37$. Fig. 9 shows the performance for individual participants in each modality and presentation condition. Fig. 10 shows performance for each verse in each condition. Eight out of the twelve participants show an audiovisual benefit for the temporally distorted presentation condition, but all the 12 participants show such benefit for speech presented in noise. The percentage of correctly identified words for the V condition was 4.1% for the temporally distorted visual speech and 6.6% for the normal speech, $F(1, 12) = 3.33$, $\mathrm{p} = 0.12$. Note that there is no difference in lip-reading abilities in the participants across Experiments 3 and 4. Two simple t-tests comparing separately performance in each of the two V conditions across experiments showed no differences (for temporally distorted V speech, $t(25) = 2.06$, $p = 0.57$; for normal V speech, $t(25) = 2.06$, $p = 0.12$).

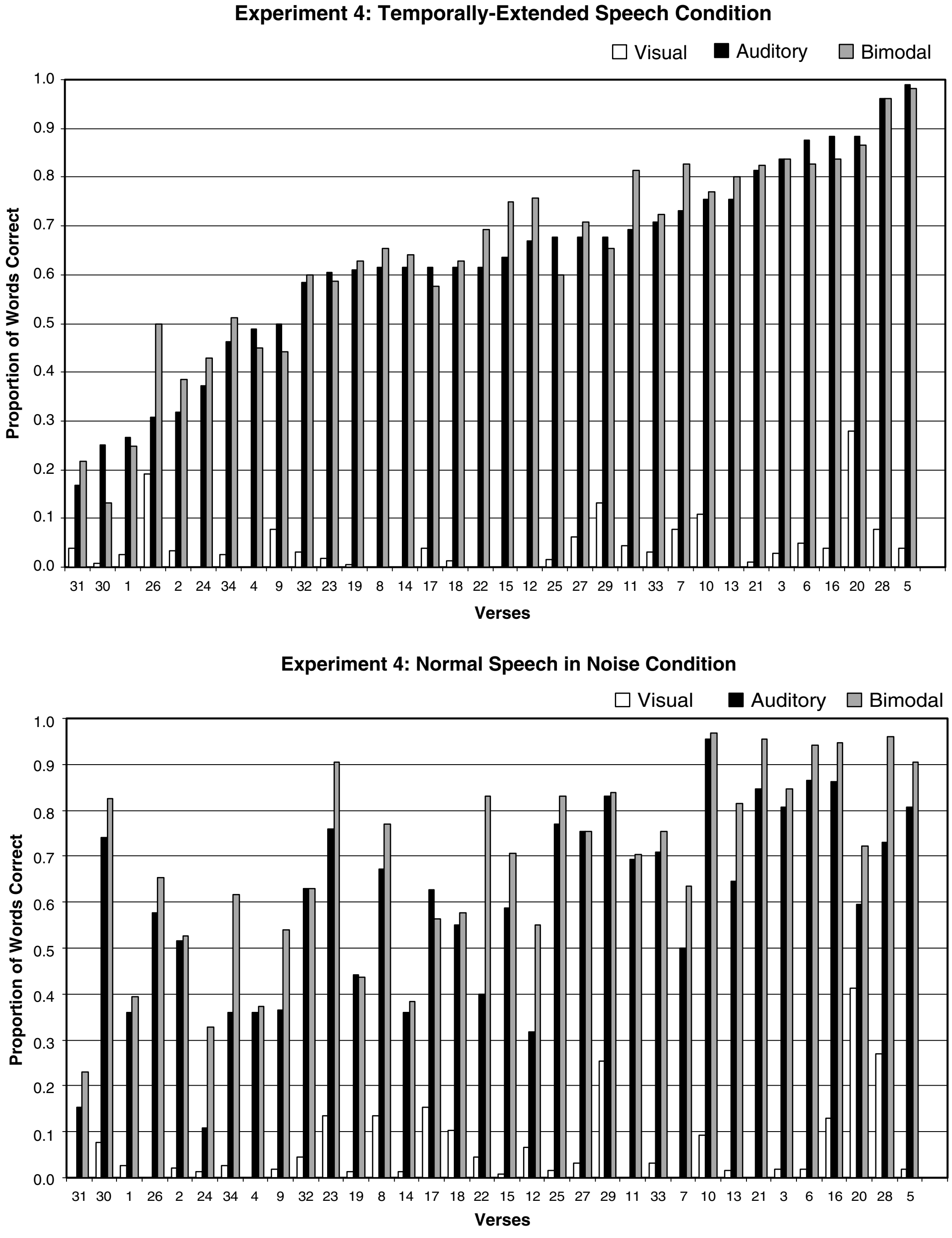

Fig. 10. The individual results for verses for the three modality conditions for the temporally distorted speech and normal-speech-in noise conditions in Experiment 4: A, V, and AV. The results are arranged from left to right in the temporally distorted speech condition according to performance on the A condition. The normal-speech-in noise condition maintains this ordering.

5.3. Discussion

Experiment 4 replicates the results from Experiments 2 and 3 in that the AV benefit was found for speech presented in noise, but not for speech that was temporally distorted. The benefit in performance in A and AV condition was smaller for temporally distorted speech than for speech presented in noise.

Given the observed variability in the relative influence of the face across the last three experiments for temporally distorted speech, we carried out an additional set of analyses including experiment as a factor. The three experiments involved exactly the same set of experimental conditions and procedures so that we simply included experiment as an additional factor. We only included data from the temporally distorted speech condition, since this condition was the same across experiments. For the speech in noise condition, different levels of noise had been added across experiments and therefore we did not compare performance in this condition across the different experiments. We restricted the analysis to an overall comparison between performance in the A condition with performance in the AV condition for temporally distorted speech. In the previous three experiments, this difference was only marginally significant. Here for the pooled data, the difference was significant ($t(38) = 2.59$, $p < 0.01$). But note that performance in the AV condition was with 62% of the words correctly recognized only 2% better than in the A condition.

Given the large difference in the two types of presentation (normal-duration speech in noise versus temporally distorted speech), one would expect to see a difference when only visible speech was presented. We therefore pooled performance in V condition not only for temporally distorted speech over Experiments 2–4, but also for the speech in noise condition. Since the noise was only added to the audio track in the A and AV conditions, the V speech presentation was the same across experiments for this condition. There was no significant difference between presentation conditions in the V condition (6.2% versus 5.4%), $t(1,38) = 1.20$, $p = 0.24$. We attribute this lack of a difference when only visible speech was presented as a floor effect in which performance was basically at chance where no difference could be observed.

6. General discussion

According to the theoretical framework of the FLMP, music perception and the perception of sung lyrics are also a type of pattern recognition and should therefore be processed in the same way as all other patterns. Previous research has shown that the perception of music and, more specifically, the recognition of sung lyrics, benefit from the addition of a visual source of information (Hidalgo-Barnes and Massaro, 2007; Vines et al., 2006). However, the audiovisual benefit in music perception has been small compared to what is usually found for spoken language. Saldaña and Rosenblum (1993) found only a small visual contribution in their musical pluck and bow study, but did not have a direct speech comparison. Scotto di Carlo and Guaitella (2004) studied the recognition of emotion in speech and in singing under auditory, visual, and audiovisual presentations. For their study, it is not possible to conclude, however, whether speech or singing gave a larger visual benefit because of significant performance differences in the A condition. Hidalgo-Barnes and Massaro (2007) showed that the recognition of words contained in a sung phrase was better when the aligned mouthing of a computer-animated talking head was presented with the auditory singing track (Hidalgo-Barnes and Massaro, 2007). However, the audiovisual benefit here was also much smaller than what is usually found for spoken language recognition. Since there was no test of these lyrics when spoken, no conclusions could be made about whether this particular set of lyrics simply did not favor a visual contribution or whether the act of singing lowered the informativeness of the visual signal.

The current study investigated why previously only a small audiovisual benefit was found for the recognition of sung lyrics (Hidalgo-Barnes and Massaro, 2007). Here, a comparison of the size of the audiovisual benefit to the one expected in speech perception was possible by comparing the recognition of the sung lyrics in the previous study with recognition of the same lyrics when spoken. In addition, we tried to overcome the limitations of the Scotto di Carlo and Guaitella study (2004) by equating the auditory performance level by adding noise to the speech.

First, we investigated the role of durational changes introduced through singing on the audiovisual intelligibility of words. The first experiment showed a similar audiovisual benefit for spoken lyrics with durations as in the original sung version compared to what was found for singing in the Hidalgo-Barnes and Massaro (2007) study. The performance on V trials was also comparable.

The three experiments compared performance for this temporally distorted presentation of the lyrics to a typically spoken version of the lyrics presented at different levels of noise. As noted by Ouni et al. (2007), the safest conclusion about differences in the size of a visual modality benefit requires a valid model of how this benefit varies with information from the auditory modality. Experiments 2–4 here found only a small audiovisual benefit for the temporally distorted version, but found substantial benefits for the lyrics spoken in noise. Although performance for V trials did not differ for these two conditions, this lack of a difference was viewed as a floor effect in which performance was basically at chance where no difference could be observed.

In summary, it appears that the temporal distortions of phonemes when sung rather than when spoken significantly attenuates the normally beneficial contribution of visible speech. Future research should evaluate real faces as well as animated ones to determine if the results can be generalized accordingly. It is also important to assess to what extent typical lyrics approximate the spoken materials that are usually used in research.

Acknowledgements

This work was supported in part by grants from the National Science Foundation (NSF Grant BCS-9905176), a grant from the Public Health Service (Grant PHS R01 DC00236), cooperative grants from the Intel Corporation and the University of California Digital Media Program (D97-04), and grants from the University of California, Santa Cruz. Baldi® is a registered trademark of Dominic W. Massaro.

Appendix A

The 34 verses used in Experiments 1 through 4. Note that 17 and 18 were actually different verses even though they had the same words.

1. by the light of lamp I sit to type
2. my notes on tap at my side
3. I don't see the sun much these days
4. a fluorescent tan covers my hide
5. how much impact shall I have this time
6. my goal today is to reach the deadline
7. I write between lines
8. I deal with fantasy
9. I report the facts
10. give them to me please
11. ham and egg salad on white bread
12. keeps me company on nights like this
13. a pack of mentholated cigarettes
14. keep my air nice and thick
15. when I write words flow like coins from a candy box
16. get out of my way I've got something to say
17. the pulse is beating louder now
18. the pulse is beating louder now
19. the cramps in my hand grow more intense with each tic tic
20. tap tap tap tap tap on the key
21. my social life is at an end
22. so it seems to be
23. why don't I trample on your lawn today
24. I'll take the sky of blue turn over old skies of grey
25. I write between the lines
26. I deal with fantasy
27. I am the press man
28. acknowledge me
29. mother always told me never stray too far from home
30. the little lady said Boy you'll never have to be alone
31. because you build with fountain pen
32. you create the memory stain
33. you are the press man
34. stand straight boy

References

Auer Jr., E.T., Kim, S., Keating, P.A., Scarborough, R.A., Alwan, A., Bernstein, L.E., 2004. Optical phonetics and visual perception of lexical and phrasal boundaries in English. J. Acoust. Soc. Am. 116, 2644.

Austin, S.F., 2007. Jaw opening in novice and experienced classically trained singer. J. Voice 21, 72–79.

Benolken, M.S., Swanson, C.E., 1990. The effect of pitch-related changes on the perception of sung vowels. J. Acoust. Soc. Am. 87, 1781–1785.

Bernstein, L.E., Eberhardt, S.P., Demorest, M.E., 1989. Single-channel vibrotactile supplements to visual perception of intonation and stress. J. Acoust. Soc. Am. 85, 397–405.

Burnham, D.K., Lau, S., Tam, H., Schoknecht, C. , 2001. Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. In: Proceedings of the AVSP, pp. 155–160.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., et al., 1997. Activation of auditory cortex during silent lipreading. Science 276, 593–596.

Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harley, F., Espesser, R., 1996. About the relationship between eyebrow movements and F0 variations. In: Proceedings of the SLP, pp. 2175–2179.

Chiappe, P., Schmuckler, M.A., 1997. Phrasing influences the recognition of melodies. Psychon. Bull. Rev. 4, 254–259.

Clarke, E., Baker-Short, C., 1987. The imitation of perceived rubato: A preliminary study. Psychol. Music 15, 58–75.

Clarke, E.F., 1985. Structure and expression in rhythmic performance. In: Howell, P., Cross, I., West, R. (Eds.), Musical structure and cognition. Academic Press, London, pp. 209–236.

Clarke, E.F., 1993. Imitating and evaluating real and transformed musical performances. Music Percept. 10, 317–341.

Cleveland, T., 1994. A clearer view of singing voice production: 25 years of progress. J. Voice 8, 18–23.

Cleveland, T., Sundberg, J., Stone, E., 2001. Long-term-average spectrum characteristics of country singers during speaking and singing. J. Voice 15, 54–60.

Cuddy, L.L., Cohen, G., Mewhort, D.J.K., 1981. Perception of structure in short melodic sequences. J. Exp. Psychol. Human Percept. Perform. 7, 869–883.

Dahl, S., Friberg, A., 2007. Visual perception of expressiveness in musicians' body movements. Music Perception 24, 433–454.

Dohen, M., Loevenbruck, H., Cathiard, M.-A., Schwartz, J.-L., 2004. Visual perception of contrastive focus in reiterant French speech. Speech Comm. 44, 155–172.

Dohen, M., Loevenbruck, H., Hill, H., 2005. A multi-measurement approach to the identification of the audiovisual facial correlates of contrastive focus in French. In: Proceedings of the AVSP, pp. 115–116.

Ellison, J.W., Massaro, D.W., 1997. Featural evaluation, integration, and judgment of facial affect. J. Exp. Psychol Human Percept Perform 23, 213–226.

Fisher, C.G., 1969. The visibility of terminal pitch contour. J. Speech Hearing Res. 12, 379–382.

Fougeron, C., Keating, P.A., 1997. Articulatory strengthening at edges of prosodic domains. J. Acoust. Soc. Am. 101, 3728–3740.

Fromkin, V., 1971. The non-anomalous nature of anomalous utterances. Language 47, 27–52.

Fry, D.B., 1955. Duration and intensity as physical correlates of linguistic stress. J. Acoust. Soc. Am. 27, 765–769.

de Gelder, B., Vroomen, J., 2000. The perception of emotion by ear and by eye. Cogn. Emotion 14, 289–311.

Granström, B., House, D., 2005. Audiovisual representation of prosody in expressive speech communication. Speech Comm. 46, 473–484.

Granström, B., House, D., Lundeberg, M., 1999. Prosodic cues in multimodal speech perception. In: Proceedings of the ICPhS, pp. 655–658.

Gregg, J.W., Scherer, R.C., 2006. Vowel intelligibility in classical singing. J. Voice 20, 198–210.

Gregory, A.H., 1978. Perception of clicks in music. Percept. Psychophys. 24, 171–174.

Hasegawa, T., Matsuki, K., Ueno, T., Maeda, Y., Matsue, Y., Konishi, Y., et al., 2004. Learned audio-visual cross-modal associations in observed piano playing activate the left planum temporale. An fMRI study. Cognit. Brain Res. 20, 510–518.

Hidalgo-Barnes, M., Massaro, D.W., 2007. Read my lips: An animated face helps communicate musical lyrics. Psychomusicology 19, 3–12.

Hnath-Chisolm, T., Kishon-Rabin, L., 1988. Tactile presentation of voice fundamental frequency as an aid to the perception of speech pattern contrasts. Ear Hearing 9, 329–334.

Hollien, H., Mendes-Schwartz, A.P., Nielsen, K., 2000. Perceptual confusions of high-pitched sung vowels. J. Voice 14, 287–298.

House, D., 2002. Perception of question intonation and facial gestures. TMH-QPSR Fonetik 44, 41–44.

House, D., Beskow, J., Granström, B., 2001. Timing and interaction of visual cues for prominence in audiovisual speech perception. In: Proceedings of the Eurospeech, pp. 387–390.

Huron, D., Ollen, J., 2003. Agogic contrast in French and English themes: Further support for Patel and Daniele (2003). Music Percept. 21, 267–271.

Jackendoff, R., Lerdahl, F., 2006. The capacity for music: what is it, and what's special about it? Cognition 100, 33–72.

Jesse, A., Vrignaud, N., Massaro, D.W., 2000/2001. The processing of information from multiple sources in simultaneous interpreting. Interpreting 5, 95–115.

Jusczyk, P.W., Krumhansl, C.L., 1993. Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. J. Exp. Psychol. Human Percept. Perform. 19, 627–640.

Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? Psychol. Bull. 129, 770–814.

Keating, P.A., Baroni, M., Mattys, S.L., Scarborough, R., Alwan, A., Auer, E.T., et al., 2003. Optical phonetics and visual perception of lexical and phrasal stress in English. In: Proceedings of the ICPhS, pp. 2071–2074.

Krumhansl, C.L., Jusczyk, P.W., 1990. Infants' perception of phrase structure in music. Psychol. Sci. 1, 70–73.

Krumhansl, C.L., Schenck, D.L., 1997. Can dance reflect the structural and expressive qualities of music? A perceptual experiment on Balanchine's choreography of Mozart's Divertimento no. 15. Musicae Sci. 1, 63–85.

Lansing, C.R., McConkie, G.W., 1999. Attention to facial regions in segmental and prosodic visual speech perception tasks. J. Speech Language Hearing Res. 42, 526–539.

Large, E.W., Palmer, C., Pollack, J.B., 1995. Reduced memory representations for music. Cognit. Sci. 19, 53–96.

Lerdahl, F., Jackendoff, R., 1983. A generative theory of tonal music. MIT Press, Cambridge, MA.

Lisker, L., 1986. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. Language Speech 29, 3–11.

Lundy, D.S., Roy, S., Casiano, R.R., Xue, J.W., Evans, J., 2000. Acoustic analysis of the singing and speaking voice in singing students. J. Voice 14, 490–493.

Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Lawrence Erlbaum Associates, Hillsdale, N.J.

Massaro, D.W., 1998. Perceiving talking faces: From speech perception to a behavioral principle. MIT Press, Cambridge, MA.

Massaro, D.W., 2002. Multimodal speech perception: a paradigm for speech science. In: Granström, B., House, D., Karlsson, I. (Eds.), Multimodality in Language and Speech Systems. Kluwer Academic Publishers, Dordrech, pp. 45–71.

Massaro, D.W., Cohen, M.M., 1999. Speech perception in perceivers with hearing loss: Synergy of multiple modalities. J. Speech Language Hearing Res. 42, 21–41.

Massaro, D.W., Egan, P.B., 1996. Perceiving affect from the voice and the face. Psychon.Bulletin Rev. 3, 215–221.

Massaro, D.W., Stork, D.G., 1998. Sensory integration and speech reading by humans and machines. American Scientist 86, 236–244.

McCrea, C.R., Morris, R.J., 2005a. Comparisons of voice onset time for trained male singers and male nonsingers during speaking and singing. J. Voice 19, 420–430.

McCrea, C.R., Morris, R.J., 2005b. The effects of fundamental frequency level on voice onset time in normal adult male speakers. J. Speech, Language, Hearing Res. 48, 1013–1026.

McCrea, C.R., Morris, R.J., 2007. Effects of vocal training and phonatory task on voice onset time. J. Voice 21, 54–63.

Miller, G., Nicely, P., 1955. An analysis of perceptual confusions among some english consonants. J. Acoust. Soc. Am. 27, 338–352.

Mixdorff, H., Charnvivit, P., Burnham, D.K. (2005). Auditory-visual perception of syllabic tones in Thai. In: Proceedings of the AVSP, pp. 3–8.

Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E., 2004. Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol. Sci. 15, 133–136.

Neuhaus, C., Knosche, T.R., Friederici, A.D., 2006. Effects of musical expertise and boundary markers on phrase perception in music. J. Cognit. Neurosci. 18, 472–493.

Nicholson, K.G., Baum, S., Kilgour, A., Koh, C.K., Munhall, K.G., Cuddy, L.L., 2003. Impaired processing of prosodic and musical patterns after right hemisphere damage. Brain Cognit. 52, 382–389.

Omori, K., Kacker, A., Carroll, L.M., Riley, W.D., Blaugrund, S.M., 1996. Singing power ratio: Quantitative evaluation of singing voice quality. J. Voice 10, 228–235.

Ouni, S., Cohen, M.M., Ishak, H., Massaro, D.W., 2007. Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. EURASIP J. Audio, Speech, Music Proc. 2007 (doi: 10.1155/2007/47891) http://www.hindawi.com/GetArticle.aspx?doi=10.1155/2007/47891>.

Palmer, C., 1989. Mapping musical thought to musical performance. J. Exp. Psychol. Human Percept. Perform. 15, 331–346.

Palmer, C., 1992. The role of interpretive preferences in music performance. In: Jones, M.R., Holleran, S. (Eds.), Cognitive Bases of Musical Communication. American Psychological Association, Washington, DC, pp. 249–262.

Palmer, C., Hutchins, S., 2006. What is musical prosody. In: Ross, B.H. (Ed.), Psychology of Learning and Motivation. Elsevier, Amsterdam, pp. 245–278.

Palmer, C., Kelly, M.H., 1992. Linguistic prosody and musical meter in song. J. Memory Lang. 31, 525–542.

Palmer, C., Krumhansl, C.L., 1987. Independent temporal and pitch structures in determination of musical phrases. J. Exp. Psychol. Human Percept. Perform. 13, 116–126.

Palmer, C., Krumhansl, C.L., 1990. Mental representations for musical meter. J. Exper. Psychology: Human Perception Perform. 16, 728–741.

Palmer, C., van de Sande, C., 1995. Range of planning in music performance. J. Exp. Psychol. Human Percept. Perform. 21, 947–962.

Patel, A.D., Daniele, J.R., 2003a. An empirical comparison of rhythm in language and music. Cognition 87, B35–B45.

Patel, A.D., Daniele, J.R., 2003b. Stress-timed vs. syllable-timed music? A comment on Huron and Ollen (2003). Music Percept. 21, 273–276.

Patel, A.D., Daniele, J.R., 2006. Comparing the rhythm and melody of speech and music: The case of British English and French. J. Acoust. Soc. Am. 119, 3034–3047.

Penel, A., Drake, C., 2004. Timing variations in music performance: musical communication, perceptual compensation, and/or motor control? Percept. Percept. Psychophys. 66, 545–562.

Primus. 1993. The Pressman. From the album Pork Soda. Interscope Records.

Repp, B.H., 1992. Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. Cognition 44, 241–281.

Repp, B.H., 1995. Detectability of duration and intensity increments in melody tones: A partial connection between music perception and performance. Percept. Psychophys. 57, 1217–1232.

Repp, B.H., 1998. Variations on a theme by Chopin: Relations between perception and production of timing in music. J. Exper. Psychology: Human Perception Perform. 24, 791–811.

Risberg, A., Lubker, J., 1978. Prosody and speech-reading. Speech Transmission Lab. Quart. Progr. Status Rep. 4, 1–16.

Rossing, T.D., Sundberg, J., Ternstroem, S., 1986. Acoustic comparison of voice use in solo and choir singing. J. Acoust. Soc. Am. 79, 1975–1981.

Saldaña, H.M., Rosenblum, L.D., 1993. Visual influences on auditory pluck and bow judgments. Percept. Psychophys. 54, 406–416.

Schmuckler, M.S., 1989. Expectation in music: Investigation of melodic and harmonic processes. Music Percept. 7, 109–150.

Scotto di Carlo, N., Germain, A., 1985. A perceptual study of the influence of pitch on the intelligibility of sung vowels. Phonetica 42, 188–197.

Scotto di Carlo, N., Guaitella, I., 2004. Facial expressions of emotion in speech and singing. Semiotica 149, 47–56.

Sloboda, J.A., 1983. The communication of musical metre in piano performance. Quart. J. Exp. Psychol. A 35, 377–396.

Sloboda, J.A., Gregory, A.H., 1980. The psychological reality of musical segments. Can. J. Psychol. 34, 274–280.

Smith, G.P., 2003. Music and mondegreens: extracting meaning from noise. ELT J. 57, 113–121.

Smith, L.A., Scott, B.L., 1980. Increasing the intelligibility of sung vowels. J. Acoust. Soc. Am. 67, 1795–1797.

Srinivasan, R.J., Massaro, D.W., 2003. Perceiving prosody from the face and voice: distinguishing statements from echoic questions in English. Language Speech 46, 1–22.

Stone, E., Cleveland, T., Sundberg, J., Prokop, J., 2003. Aerodynamic and acoustical measures of speech, operatic, and Broadway vocal styles in a professional female singer. J. Voice 17, 283–297.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Summerfield, Q., 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, B., Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lip-reading. Erlbaum, London, pp. 3–51.

Sundberg, J., 1974. Articulatory interpretation of the singing formant. J. Acoust. Soc. Am. 55, 838–844.

Sundberg, J., 1982. Perception of singing. In: Deutsch, D. (Ed.), The Psychology of Music. Academic Press, New York, pp. 59–98.

Sundberg, J., 2003. Research on the singing voice in retrospect. TMH-QPSR Speech Music Hearing 45, 11–22.

Sundberg, J., Skoog, J., 1997. Dependence of jaw opening on pitch and vowel in singers. J. Voice 11, 301–306.

Swerts, M., Krahmer, E. (2004). Congruent and incongruent audiovisual cues to prominence. In: Proceedings of the Speech Prosody.

Swerts, M., Krahmer, E., 2005. Audiovisual prosody and feeling of knowing. J. Memory Lang. 53, 81–94.

Tan, N., Aiello, R., Bever, T.G., 1981. Harmonic structure as a determinant of melodic organization. Memory Cognition 9, 533–539.

Thompson, D.M., 1934. On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues. J. Gen. Psychol. 11, 160–172.

Titze, I.R., Sundberg, J., 1992. Vocal intensity in speakers and singers. J. Acoust. Soc. Am. 91, 2936–2946.

Todd, N.P.M., 1995. The kinematics of music expression. J. Acoust. Soc. Am. 97, 1940–1949.

Trainor, L.J., Adams, B., 2000. Infants' and adults' use of duration and intensity cues in the segmentation of tone patterns. Percept. Psychophys. 62, 333–340.

Vatakis, A., Spence, C., 2006. Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. Neurosci. Lett. 393, 40–44.

Vines, B.W., Krumhansl, C.L., Wanderley, M.M., Levitin, D.J., 2006. Cross-modal interactions in the perception of musical performance. Cognition 101, 80–113.

Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E., 2002. Linking facial animation, head motion and speech acoustics. J. Phonetics 30, 555–568.