# Visual speaker gender affects vowel identification in Danish

*Charlotte Larsen & John Tøndering*
*Department of Scandinavian Studies and Linguistics, University of Copenhagen, Denmark*

## Abstract

*The experiment examined the effect of visual speaker gender on the vowel perception of 20 native Danish-speaking subjects. Auditory stimuli consisting of a continuum between /muːlə/ 'muzzle' and /moːlə/ 'pier' generated using TANDEM-STRAIGHT matched with video clips of a female and a male speaker were used to determine whether visual speaker gender affected Danish listeners similarly to American English-speaking listeners tested in a similar way.*

## Introduction

The purpose of the experiment reported here is to determine whether visual information about speaker gender affects Danish listeners' vowel perception. Participants were presented with audiovisual stimuli consisting of combinations of a woman's and a man's face with two auditory continua with vowel qualities between /muːlə/ 'muzzle' and /moːlə/ 'pier', one with a woman's voice, the other with a man's. Based on similar experiments conducted with American English speakers and listeners, the participants were expected to identify more steps on both auditory continua as /muːlə/ when they were paired with the female face than when they were paired with the male.

The existence of such an effect of visual gender on the vowel perception of Danish listeners would not only provide information on listeners' expectations regarding women's and men's speech, it would also contribute to theories of speech perception, and in particular to theories of speaker normalization.

### Speaker normalization

It is widely accepted that most men's voices sound different from most women's. This is partly due to differences in average size and shape of the vocal tract – e.g. women's vocal tracts tend to be shorter than men's, producing vowels with higher average formant frequencies (Ladefoged & Broadbent, 1957). Differences may also result from factors not directly linked to physical differences, as evidenced by the fact that listeners are able to tell girls' voices apart from boys' before the children are old enough to have developed the physiological differences which might otherwise account for voice differences (Perry, Ohde & Ashmead, 2001). This suggests that social or cultural factors may also be involved.

For a variety of reasons, then, sounds which listeners have no trouble categorizing as instances of the same phoneme are acoustically quite variable depending on the gender of the speaker. Theories of speech perception need to be able to account for this through an explanation of the phenomenon known as speaker normalization, the process by which listeners fit input from individual speakers to phoneme categories available in their language.

Theories of speaker normalization have traditionally focused on physical differences between speakers, for example the theory, a version of which is put forth by Potter and Steinberg (1950), that sets of vowels produced by different speakers have about the same *relative* distribution in acoustical or auditory space. Another theory holds that listeners construct a mental model of a speaker's vocal tract, allowing them to correct for the effect of its size and shape and extract a set of absolute formant frequencies common to all speakers of a particular dialect independent of what Joos (1948) terms 'PERSONAL ERROR' (1948: 6).

Neither theory accounts for variation caused by non-physiological factors, however, and the contribution of visual cues to speech perception is ignored entirely.

### Visual integration in speech perception

Perhaps the most famous example of how visual information affects speech perception is the McGurk-effect (McGurk & MacDonald, 1976) which demonstrated how listeners could be made to perceive a third sound by presenting them with audiovisual stimuli with visual articulatory information pointing to one sound, auditory information to another.

Later experiments have shown a similar effect of articulatory information in vowel perception, e.g. Traunmüller and Öhrström (2006).

However, there is evidence that information about place or manner of articulation is not the only type of visual information relevant to vowel perception. In a series of experiments, Johnson, Strand and D'Imperio (1999) found that American English listeners were likely to perceive more steps on a continuum of auditory stimuli comprising a continuum between /hood/ and /hud/ as /hood/ when they were paired with a video clip of a woman speaking than when the speaker was visually male. An effect was found not just of visual gender but also of the degree of gender stereotypicality of different voices and faces as judged by another group of participants. This last finding in particular would be difficult to explain under a theory of speaker normalization which only concerns itself with average physiological gender differences.

Based on these findings, which suggested that more than one kind of visual information was integrated during speech perception, the authors advocated a theory of speaker normalization which includes 'abstract, subjective talker representations' (ibid; 1999:380).

The experiment described in this paper is based on one of the experiments presented in Johnson, Strand and D'Imperio (1999), aiming to discover whether a similar effect of visual gender can be shown for Danish speakers and listeners.

## Method

The stimuli were produced using sound and video recordings (head and shoulders) of a 31 year old woman and a 34 year old man pronouncing the words /mu:lə/ and /mo:lə/. Video was recorded in QuickTime format in the resolution 1920×1080 using a JVC GY-MH100 camera, while sound was recorded in wav format, 16 bit, 44,100 Hz stereo using an Olympus LS10 digital recorder.

Several repetitions of each word were produced, and video and sound clips chosen for further manipulation did not come from the same instance in the original recording, removing the risk of some finished audiovisual stimuli seeming better synchronized than others.

In order to avoid a learning effect relating a particular intonation to a particular vowel quality, the four sound clips were manipulated in Praat (v. 5.3.03) to keep F0 constant throughout each clip and identical for the two clips produced by the same speaker, at 220 Hz for the female voice, 133 Hz for the male voice, these frequencies being the averages of the average pitch values for each pair of clips.

The two pairs of words were manipulated using TANDEM-STRAIGHT, a speech manipulation program which allows auditory morphing based on source-filter analysis of recorded speech (Kawahara et al., 2009). For each speaker, a continuum was generated with nine steps between [mo:lə] and [mu:lə] which will be referred to as auditory stimulus 1–9, 1 being 100% [mo:lə], 9 being 0% [mo:lə], that is, 100% [mu:lə]. As the purpose of the experiment was primarily to reveal the difference in perceptual phoneme boundary of one set of stimuli compared to others, and not to provide an absolute value of, e.g., frequency, no perceptually motivated scale was used to determine the degree of morphing; the nine steps were simply morphed with equal percentual intervals so that stimulus 5 equals 50% morphing between [mo:lə] and [mu:lə].

Likewise, for each stimulus all parameters which TANDEM-STRAIGHT manipulates were set to the same degree of morphing, except 'Time' which was set to 50% morphing for all stimuli in order to avoid differences in synchronization between sound and video. This approach yielded a continuum of auditory stimuli which are morphed in more dimensions than, for example, a continuum of synthetically generated vowels inserted in the desired context, but it also means that it is impossible to determine exactly which acoustical features were relevant for participants' perception of vowel quality in the finished stimuli.

Using Final Cut Pro 7, each of the 18 auditory stimuli (two voices × nine steps on the vowel quality continuum) was paired with each of the four visual stimuli (two faces × two visual pronunciations, /mo:lə/ and /mu:lə/), creating a total of 72 different audiovisual stimuli. In order to avoid a sequence effect, stimuli were administered using one of four randomized lists.

Twenty-three linguistics students from University of Copenhagen participated in the experiment. Of these, 20 had Danish as their native language while three had Faroese or Faroese and Danish. Only the replies of participants with Danish as their (only) native language were included in the study. Of these 20, 12 were female, 8 were male. Average birth year was 1988, median birth year 1990, and participants' regional backgrounds were mixed, with 75% having been raised on Zealand.

Participants were tested individually in a quiet room using a laptop computer and a set of headphones. As it was vital for participants to keep their eyes on the screen while the stimuli were played, they were instructed to answer verbally, and in order to avoid any effect of participants themselves pronouncing the stimulus words between stimuli, the replies were given in the form of the numbers 'one' and 'two' rather than the words themselves.

## Results

Overall, the results of the experiment showed the expected effect of visual speaker gender on vowel perception. 56.1% of stimuli with the visually female speaker were perceived by participants as /mu:lə/ while the same was true for only 50.4% of stimuli with the visually male speaker. A chi squared test showed this difference to be significant ($p < 0.01$).

There were, however, substantial differences in the way the effect manifested itself in different sets of stimuli – or failed to show up at all.

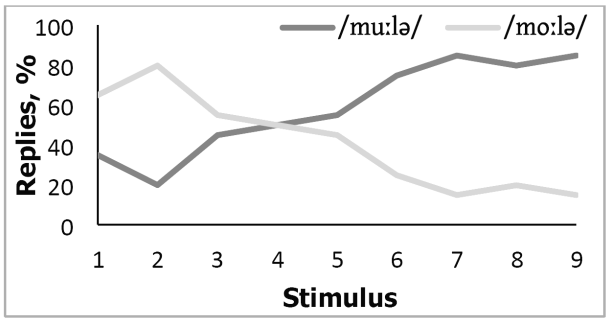


*Figure 1. Replies for the set of stimuli with visual /mo:lə/-pronunciation+female voice+female face.*

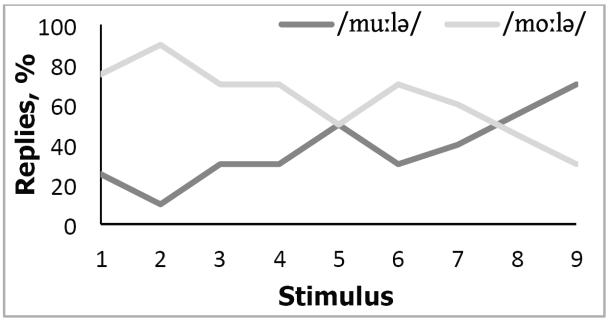


*Figure 2. Replies for the set of stimuli with visual /mo:lə/-pronunciation+female voice+male face.*

*Figures 1–2* are examples of the visual representation of the results, showing the answers for two sets of stimuli which differ only in the variable visual gender. The point where answers are split evenly between /mu:lə/ and /mo:lə/ was chosen as a convenient numerical measure for further analysis, hereafter named *the perceptual crossover point*. It should be mentioned that some sets had two potential crossover points, as seen in *Figure 2*. Here, 50% of participants perceived stimulus 5 as /mu:lə/, however, as all stimuli 1–7 with the exception of stimulus 5 were perceived as /mo:lə/ by more than 50% of participants, 7.67 is considered the actual perceptual crossover point.

*Table 1. Perceptual crossover points for all combinations of visual and auditory gender and visual pronunciation.*

| | Visual /mo:lə/ | | Visual /mu:lə/ | |
|---|---|---|---|---|
| | Female voice | Male voice | Female voice | Male voice |
| Female face | 4 | 5.63 | 4.14 | 3.6 |
| Male face | 7.67 | 7.25 | 2.2 | 4.64 |

As *Table 1* shows, the expected effect – a perceptual crossover point closer to 1 (= more /mu:lə/-answers) for stimuli sets with the female face than with the male – is seen for three out of four combinations of auditory gender and visual pronunciation. However, for the set with a male voice and visual /mu:lə/-pronunciation, the number of /mu:lə/-replies across all stimuli is actually close to being the same for both visual genders, 56.1% for the visually male speaker, 55% for the visually female one – a small difference in the opposite direction of the one predicted by the hypothesis, despite the perceptual crossover point for the set with the visually female speaker being closer to 1 as predicted.

A possible explanation for this discrepancy is found in the fact that there is considerably less agreement about the classification of stimuli for the set showing a female face with a male voice than for the set with matched visual and auditory gender. Stimulus 2–6 in the gender-mismatched set were each identified by less than 70% of participants as being either /mu:lə/ or /mo:lə/, and no stimulus in the set was classified the same by 90% of participants. Generally, there was less agreement about the sets with mismatch between visual and auditory gender than about the ones with matched genders, possibly because participants were aware of and distracted by the discrepancy.

The variables auditory gender and visual pronunciation were also found to have the expected effect on vowel perception, that is, listeners identified significantly more stimuli as /muːlə/ when they heard the female voice or saw the speaker of either gender pronouncing this word.

## Discussion and conclusion

The experiment demonstrated the integration during vowel perception of two distinct kinds of visual information: articulatory information, the effect of which was expected based on the findings of e.g. Traunmüller and Öhrström (2006), and visual information about speaker gender. The effect of visual speaker gender was similar to the one found by Johnson, Strand and D'Imperio (1999), so the results appear to support their view of speaker normalization as based on several different kinds of information and partly dependant on listeners' representations of speakers, not only with respect to the size of their vocal tract.

While the findings certainly support this view, when taken alone, they are not strictly incompatible with a theory of normalization based on individual physical differences – e.g. listeners may simply have noted that the male speaker was larger than the female. An effect of the perceived visual gender stereotypicality of speakers independent of speaker size would disprove this alternative explanation, and this will be the focus of further research.

Regarding the methodology of the experiment, it should be mentioned that, surprisingly, out of all audiovisual stimuli, only one was identified by all twenty participants as the same phoneme. To our ears, the end points of each manipulated auditory continuum were all clearly identifiable as the word they were 'meant' to represent when heard in isolation, suggesting that the ambiguity arose from the combination of auditory and visual stimuli, but as this was not verified by a separate test, we cannot rule out the possibility that the process used for auditory morphing in itself introduced an unintended perceptual ambiguity.

Furthermore, as only four video clips were used, each representing a combination of gender and pronunciation, variation between clips, such as overarticulation on one clip, may have seriously impacted results, masking the effect of visual gender. Analysis of results broken down by visual stimulus suggests this may well have been the case for the two sets of stimuli which did not show the expected effect, but further research would be necessary to determine whether this was in fact the case, as well as to determine the exact relation between the method of auditory manipulation and the perceptual ambiguity discussed above.

Finally, as the participants in this study were not selected to be representative, and there is a strong possibility that participant variables such as age, gender and regional background affect the outcome, the findings cannot be said to apply to Danish listeners in general. The mere fact that an effect was shown for this particular group does however demonstrate that the effect of visual gender on vowel perception is not unique to the American English-speaking populations examined by Johnson, Strand and D'Imperio (1999) and others, and underscores the need for theories of speaker normalization to take into account not just physical differences between speakers of different genders but also, for example, listener expectations of how women and men are 'supposed' to speak.

## References

Johnson, K., E. A. Strand & M. D'Imperio. 1999. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27:359−384.

Joos, M. A. 1948. Acoustic phonetics. *Language* Supplement 24(2):1–136.

Kawahara, H., T. Toru, M. Takahashi, M. Morise & H. Banno. 2009. Development of exploratory research tools based on TANDEM-STRAIGHT. *Proc. APSIPA*, Sapporo, 111–120.

Ladefoged, P. & D. D. Broadbent. 1957. Information Conveyed by Vowels. *Journal of the Acoustical Society of America* 29(1):98–104.

McGurk, H. & J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264:746–748.

Perry, T. L., R. N. Ohde & D. H. Ashmead. 2001. The acoustic bases for gender identification from children's voices. *Journal of the Acoustical Society of America* 109(6):2988–2998.

Potter, R. K & J. C. Steinberg. 1950. Toward the Specification of Speech. *Journal of the Acoustical Society of America* 22(6):807–820.

Traunmüller, H. & N. Öhrström. 2007. Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics* 35:244–258.

# Proceedings of Fonetik 2013

## The XXVI$^{th}$ Annual Phonetics Meeting 12–13 June 2013, Linköping University Linköping, Sweden

Studies in Language and Culture
no. 21

Robert Eklund, editor


Linköping University

Conference website: www.liu.se/ikk/fonetik2013
Proceedings also available at: http://roberteklund.info/conferences/fonetik2013

Cover design and photographs by Robert Eklund
Photo of Claes-Christian Elert taken by Eva Strangert on the occasion of his 80th birthday

Proceedings of Fonetik 2013, the XXVI$^{th}$ Swedish Phonetics Conference
held at Linköping University, 12–13 June 2013
Studies in Language and Culture, no. 21
Editor: Robert Eklund
Department of Culture and Communication
Linköping University
SE-581 83 Linköping, Sweden

ISBN 978-91-7519-582-7
eISBN 978-91-7519-579-7
ISSN 1403-2570



Printed by LiU-Tryck, Linköping, Sweden, 2013