SPEECH PERCEPTION AT ITS BEST: EXTRACTING LINGUISTIC INFORMATION FROM ACOUSTICALLY UNDERSPECIFIED INPUT. THE CASE OF SINGING

Andrea Deme

Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest
Eötvös Loránd University, Budapest
deme.andrea@nytud.mta.com

ABSTRACT

High-pitched sung vowels are "underspecified" due to i) the tuning of the F1 to the f0 accompanying pitch-raising, and ii) the wide harmonic spacing of the voice source resulting in the undersampling of the vocal tract transfer function. Therefore, sung vowel intelligibility is expected to decrease as the f0 increases. On the basis of the literature of speech perception it is often suggested that sung vowels are better perceived if uttered in CVC context (than in isolation) even at high f0, but the results for singing are contradictory. In the present study we further investigate this question. We compare vowel identification in sense and nonsense CVC sequences and show that the positive effect of the context disappears if the number of legal choices is similar in both conditions, meaning that any positive effect of the CVC context may only stem from the smaller number of possible responses, i.e. higher probabilities.

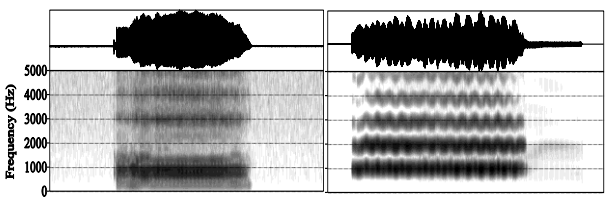
Keywords: high-pitched sung vowel, consonant environment, vowel identification

1. INTRODUCTION

High-pitched sung vowels are underspecified for two reasons. First, singers tune the F1 to (or slightly above) the f0 if the f0 is raised, and thus they compensate for the loss of acoustic energy and timbre inhomogeneity resulting from f0 exceeding the value of F1 of the vowels. Second, the wide harmonic spacing of the voice source results in undersampling of the vocal tract transfer function, i.e. in the low resolution of the articulated sound (see e.g. [2] and Figure 1). By tuning F1 to (or slightly above) the f0, singers systematically change the articulatory-acoustic vowel target as the f0 increases [2], while the spectrum also gets denser and the resolution of the vowel becomes lower. As a result, the decrease of the intelligibility of sung vowels is also expected and was observed in several studies [3, 4, 5].

However, in consistence with the literature of speech perception [7] it can also be assumed that vowel identity can be preserved to some extent by its consonantal context even at high f0. The two prior studies that investigated this assumption found contradictory results: while [6] claimed to prove it in sense words, [3] showed no positive effect of consonantal context in nonsense sequences. It should be noted, that due to the sense-nonsense opposition, these two studies investigated two essentially different aspects of the question. While in [6] the task of the listeners' was to identify meaningful words (thus both bottom-up and top-down perceptual processes were both activated), in [3] the task was to identify vowels in nonsense carrier sequences (where only bottom-up analysis is available). Moreover, in both studies vowel identification in CVC sequences was compared to vowel identification in isolation which task is evidently more similar to the vowel identification task in nonsense words (as in both cases only the acoustic cues are available for the perceptual system). Finally, (as opposed to [3]) [6] compared vowel identification in conditions that were not equally balanced in number: the number of possible responses in CVC condition, i.e. the number of sense words, was only four compared to the isolated vowel condition in which at least 12 competing responses were allowed (the Methods section of [6] does not clarify this issue sufficiently). The contradictory results obtained in [3] and [6] may only be in apparent contradiction, which becomes obvious if reformulated as follows: as [6] provided evidence that **vowel identification in meaningful words** might be more effective than that in isolation, [3] showed that **consonantal context as a source of acoustic cues** cannot take a positive effect on the identification of the intended sung vowel quality by itself.

Figure 1: The waveform and the spectrum of the vowel /a:/ in speech (~200 Hz) and at a high f0 (b5 = 988 Hz) sung by a professional soprano.



Hence, the question arises whether the positive effect of sense or meaning persists if the number of choices is balanced in the sense and nonsense conditions. On this basis, the aim of the present study is to compare vowel identification in sense and nonsense words, to clarify how the consonantal context affects vowel perception in singing. Based on the findings of [3] and [6] it is hypothesized that vowel identification is more effective in sense words than in nonsense CVC sequences.

2. PARTICIPANTS, MATERIAL, METHODS

To test the hypothesis a vowel identification test was carried out in a group of 20 healthy Hungarian adults (32-year-olds on average).

For the perception test two sets of stimuli were created and recorded in a sound-treated room sung by a professional soprano. One set served as the *nonsense* condition, while the other one served as the *sense* condition. Both conditions constituted of CVC-structured sequences. To maximize the number of options (thus to reduce the probability of one vowel in the test) and to control for the coarticulatory effects caused by the place of articulation of the consonant, the Hungarian vowels /ɒ aː ɛ iː øː yː/ were selected for testing in alveolar context. In *sense* condition the consonantal frame consisted of /sVr/ resulting in six meaningful Hungarian words (e.g. /saːr/ 'stalk'). In *nonsense* condition the context /dVr/ was used, that resulted in six non-meaningful syllables (e.g. /daːr/). The test sequences were sung at six different fundamental frequencies from 178 Hz to 988 Hz (covering the soprano range) and were also uttered in speech (at approx. 199 Hz) by the soprano.

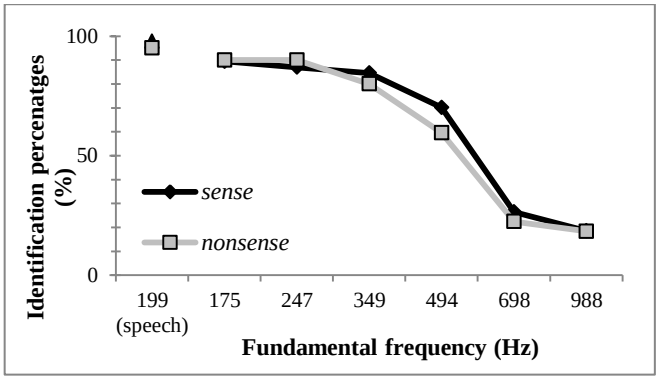
The perception test was controlled by a Praat script [1]. The 84 test stimuli were presented to each participant twice (with 42 filler sequences) in a randomized order in two separate sessions for the two conditions (starting with the *sense* condition). In both sessions, the task of the participants was to choose the vowel they heard from the 9 Hungarian vowels /ɒ aː ɛ eː iː oː øː uː yː/ presented on the computer screen. Consequently, in the first session participants were also presented responses (visually) that were not legal in the *sense* condition (/eː/). As /oː/ and /uː/ would combine into sense Hungarian words even in the *nonsense* context (as /doːr/ 'dorian scale' and /duːr/ 'major scale'), these vowels were not investigated in the study (but were used as fillers). In the first session the participants were informed that they were going to hear Hungarian words, while in the second session they were warned that sense and nonsense words may both be presented, thus they must focus only on the quality of the vowel in question. In each trial, the consonantal context was also presented to the participants visually at the top of the screen (to maximize the possible support provided by coarticulatory effects), as well as the list of all legal responses (i.e. all possible words) in the *sense* condition. The sound stimuli were presented through headphones.

3. RESULTS

3.1. Overall vowel identification tendencies as a function of f0 and the mode of production

Overall rates of vowel identification in *sense* and *nonsense* conditions are shown in Figure 2 where each point represents the percentage of correct responses out of the 240 stimuli presented at the particular f0. (Please note that as vowel targets are assumed to change as a function of pitch, "correct" responses and "errors" refer to the match and mismatch between the intended vowel quality and the response). It is clearly seen that vowel identification tendencies are basically identical in sense words and nonsense CVC sequences (no significant difference according to a one-way ANOVA). The highest percentages of identification are observable in speech mode in both conditions, and the participants' performance decreases rapidly after the f0 of 494 Hz (the musical note b4 or *b'*).

Figure 2: Overall rates of vowel identification in the *sense* and *nonsense* conditions.



3.2. Vowel identification as a function of f0

Figure 3 summarizes the results on each of the six target vowels used in the two conditions. Generally, no significant difference was found between the two conditions (according to a repeated measures ANOVA).

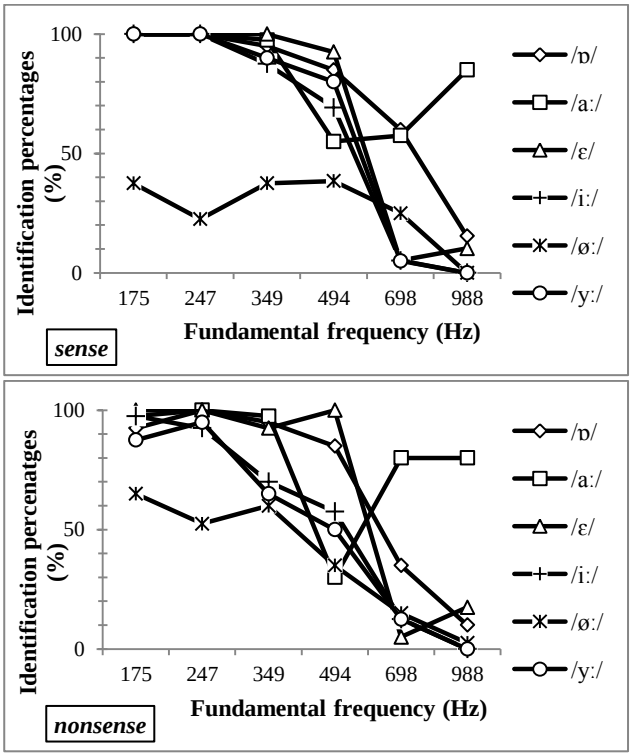
Although participants had a slightly higher chance of choosing the correct vowel by chance in the *sense* condition (due to the smaller number of alternatives) than in the *nonsense* condition, participants' performance was still lower in the case of /øː/

in the *sense* condition even at low f0s (below the apparently critical 698 Hz). Furthermore, /øː/ was identified in only 87.5% in speech, whereas all the others vowels were identified in 100% (not presented in Figure 3).

The highest rate of identification accuracy is observable in the case of /aː/ in both conditions. However, a sudden dip occurs at 494 Hz, again, simultaneously in both conditions.

With respect to the vowel identification below 698 Hz (where the performance level was generally higher), the results on close front vowels /iː yː/ are higher in the *sense* condition than in the *nonsense* condition, thus the tendencies of /iː yː/ and the more open /ɒ ɛ/ in the *sense* condition are more similar than in the *nonsense* condition. In other words, the close /iː yː/ were perceived at a similar rate as the more open /ɒ ɛ/, if the close vowels were uttered in the Hungarian words /siːr/ (‘Syrian’) and /syːr/ (‘to filter’ or a traditional Hungarian couched jacket), than in nonsense sequences. However, participants again had a higher chance of choosing the correct response by chance in the *sense* condition.

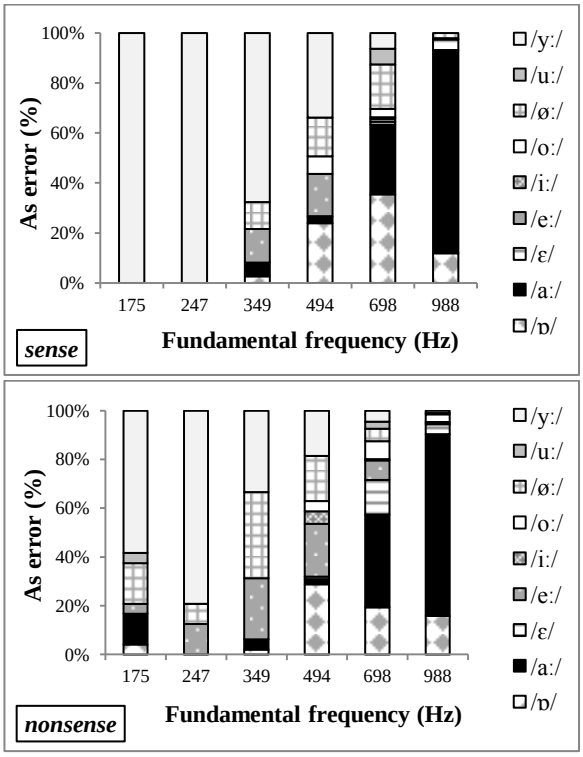
Figure 3: Identification of vowels in the *sense* condition (upper panel), and in the *nonsense* condition (bottom panel).



3.3. Error tendencies of vowel identification

Figure 4 shows the error tendencies as a function of f0. For example, 100% of /yː/ responses in the *sense* condition (upper panel) at the lowest f0 (175 Hz) indicates that 100% of all the errors (mismatches) occurring at 175 Hz in *sense* condition were misinterpretations of any target sound as /yː/.

Figure 4: Vowels occurring as errors among all observed errors as a function of f0 (upper panel: *sense* condition, bottom panel: *nonsense* condition).



In the case of the lowest two f0s (175 Hz and 247 Hz) /yː/ was the most frequent error in both conditions as a result of the high ratio of misinterpreted /øː/ sounds. According to Figure 4, above 494 Hz (where identification rates decrease rapidly) the main tendencies of vowel misidentifications consist of mismatches for /aː/ /ɒ/ and /øː/ in both conditions and with /eː/ mainly in the *nonsense* condition. It should be emphasized again that the identification of the vowels /ɒ aː ɛ iː øː yː/ was investigated in the study, but the vowels /ɒ aː ɛ eː iː oː øː uː yː/ were all possible candidates the listener could choose in both conditions during the listening experiment. This is particularly important, because Figure 4 shows that although listeners were warned not to choose illegal answers in the *sense* condition, they could not help doing so: they chose the vowel /eː/ as a response to the (intended) /iː/ target even at and below the apparently critical f0, 494 Hz. Interestingly, despite the fact that /soːr/ and /suːr/ are sense Hungarian words (but were considered illegal in the present study, see section 2 for further discussion), but /seːr/ is nonsense, the illegal answers /oː uː/ were much less frequently chosen than /eː/. The increased percentage of

/yː/ mismatches is mainly the result of misinterpreted /øː/ sounds at all f0s in both conditions.

4. DISCUSSION AND CONCLUSIONS

The results did not verify our hypothesis: the identification of high-pitched sung vowels did not show to be more effective in sense words than in nonsense words, even if the chances of choosing the correct response were slightly higher in the *sense* condition. However, moderately higher identification percentages of certain vowels in sense words at low f0s were also found.

The finding that /iː/ was identified according to the intention of the singer at a moderately higher percentage in the *nonsense* condition, may be to some extent explained on the basis of previous research. According to e.g. [8] and [2] the modification of close vowels accompanying pitch raising results in the increase of F1, which in the case of close vowels (but not in mid vowels) also leads to the percept of vowels with higher F1 than the intended vowel quality (see [3]). Based on this, it can be implied that in the present study listeners might have perceived /eː/ instead of /iː/ below 988 Hz at a similar rate in both conditions. However, the ratio of /eː/ responses was different, as /eː/ was an illegal response in the *sense* condition and thus it was more frequently repressed than in the *nonsense* condition. Instead, participants were forced to choose /iː/. This claim is also supported by the high ratio of /eː/ as an illegal response (in the *sense* condition), which reflects that the perceptual category shift was so clear that the participants could not help choosing the illegal /eː/ response.

The high perceptibility level of /aː/ and the mismatches for /ɒ/ and /aː/ near 1 kHz are also in line with previous findings that provide evidence for the acoustic stability of the vowels with high F1 through the soprano range and for the high number of mismatches for vowels with high F1 at about 1 kHz [3, 4, 5]. The data on /øː/ (reflecting that it was perceived at a low rate even in speech) suggests that this particular vowel of the singer participant was "uniquely pronounced" and was realized, to say, further away from the prototype of the perceptual category of /øː/ of the listeners at each f0.

The present findings lead to two important conclusions. First, the articulatory–acoustic vowel targets change in singing as the f0 increases, thus any positive effect attributed to the consonantal context should also be considered with taking this factor into account: if vowel targets change, a positive effect provided by the neighbouring consonants (i.e. formant transitions) may only cue the modified, but not the intended vowel quality. The lack of clear shifts from one perceived vowel category to another accompanying pitch raising, however, questions the assumption of any positive effect of the formant transitions.

Second, the findings emphasize the importance of a differentiation within the concept of **consonantal environment** (in singing), i.e. to differentiate between i) the role of **formant transitions** as acoustic cues that aid the (bottom-up) processes of speech perception, and ii) the role of **sense or meaning** that enhance the efficacy of vowel identification through the involvement of top-down processes and word/vowel probabilities.

i) The tendencies observed in the nonsense sequences suggest that the amount of the acoustic information, which would lead the listener to perceive the intended vowel quality, decreases in the speech signal with pitch raising. This means that 1. vowel targets change along with formant transitions with the increase of the f0, and 2. the undersampling effect (caused by high f0, see [2]) affects vowels, as well as the (sonorant) formant transitions resulting in underspecification of the vowels when sung at high f0. Therefore, it can be declared that formant transitions do not have a positive effect on the identification of (high-pitched) sung vowels.

ii) If we interpret our findings in the light of the results of [3] and [6], however, it can also be concluded that informational content, sense or meaning may play a role in the identification of sung vowels, but only through the number of possible items (vowels or words) to be identified, i.e. through probability. While [6] provided evidence that a smaller number of sense words are better distinguished than a larger number of isolated vowels, [3] showed that nonsense words and vowels in isolation are perceived in a similar ratio, if the responses are assigned to similar probabilities in the tasks. As the present study showed that sense or meaning does not have a positive effect on vowel identification if the number of choices is balanced, it is concluded that the effect found in [3] may only be present if the number of possible words is small restricting the identification task.

To summarize, while consonantal environment does not aid lower level speech perception processes in the identification of the intended quality of high-pitched sung vowels by coarticulatory cues (extractable from the signal), i.e. formant transitions, it aids higher level perceptual processes by restricting the number of possible answers and hence increasing their probability. Thus, we may say, the only factors vowel identification in singing is dependent on are **context**, i.e. text- and sentence-embedding, and **phonology**, i.e. phonological neighborhood density effects.

5. REFERENCES

[1] Boersma P, Weenink D. 2009. *Praat: Doing phonetics by computer*. Computer program, http://www.praat.org

[2] Deme, A. 2014. Formant strategies of professional female singers at high fundamental frequencies. *Proc. 10th ISSP* Cologne, 90–93.

[3] Deme, A. 2014. Intelligibility of sung vowels: The effect of consonantal context and the onset of voicing. *J. Voice* 28, 19–25.

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

[5] Scotto di Carlo N, Germain A. 1985. A perceptual study of the influence of pitch on the intelligibility of sung vowels. *Phonetica* 42, 188–97.

[6] Smith L, Scott B. 1980. Increasing the intelligibility of sung vowels. *J. Acoust. Soc. Am*. 6, 1795–1797.

[7] Strange W., Verbrugge R. R. 1976. Consonant environment specifies vowel identity. *J. Acoust. Soc. Am*. 60, 213–224.

[8] Sundberg, J. 1975. Formant technique in a professional female singer. *Acta Acustica united with Acustica* 32, 89–96.