The Identification of High-pitched Sung Vowels in Sense and Nonsense Words by Professional Singers and Untrained Listeners

Andrea Deme, *Budapest, Hungary*

Summary: High-pitched sung vowels may be considered phonetically "underspecified" because of (i) the tuning of the F_1 to the f_0 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 f_0 increases. Based on the literature of speech perception, it is often suggested that sung vowels are better perceived if uttered in consonantal (CVC) context than in isolation even at high f_0. The results for singing, however, 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 in a perception test 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., from higher probabilities. Additionally, it is also tested whether the training in production (i.e., singing training) may also lead to a perceptual advantage of the singers over nonsingers in the identification of high-pitched sung vowels. The results show no advantage of this kind.

Key Words: High-pitched singing–Sung vowel–Vowel identification–Consonant environment–Trained listener.

INTRODUCTION

Vowels sung by sopranos at high fundamental frequencies undergo significant changes in production; that is, both the articulation and the acoustics of high-pitched sung vowels are different from those of their spoken realizations. These changes stem from two main sources. The first source is the set of requirements of the Western operatic (hereafter, operatic) tradition on vocal timbre that demands the singing voice to be homogenous in timbre and loud without any further amplification throughout the whole pitch range. The other source is the changing interrelation of the fundamental frequency (f_0) and its harmonic partials accompanying pitch-raising.

Traditionally, sopranos of the operatic style are required to sing very loudly so that they can be easily heard through a large orchestra, while their vocal timbre remains homogenous.[1] However, this effort is largely hindered by the conflict of the production of distinct vowel qualities and the f_0-raising required by the musical score. As Sundberg[2] points out, the sopranos must relatively often sing well above the frequency region of the first formant (F_1), which, in the case of the unmodified articulation of the vowels, would result in several serious acoustic consequences. If the f_0 exceeds the F_1, the voice becomes weak and changes in timbre abruptly. As mentioned, however, both of these effects are to be avoided in singing. So what might be the solution sopranos employ to avoid undesirable changes of the vocal quality?

Sundberg suggests that at high f_0, sopranos must modify the articulatory configuration of vowels to gain loudness and homogenous timbre through the whole pitch range used in their singing practice. The most prevalent assumption on the nature of the modification of vowel production accompanying pitch raising is that in those cases where sopranos sing above the frequency region of the F_1 of spoken vowels, they tune the F_1 to the frequency of the f_0, or slightly above (this strategy is hereafter referred to as $F_1 : f_0$ for short). In spite of the difficulties of formant frequency estimation at high f_0 (for further explanation, see the description of the undersampling effect below), this suggestion is supported by a respectable amount of empirical evidence (see Sundberg,[3] Johansson et al,[4] Hertegård and Gauffin,[5] Joliveau et al,[6] Garnier et al,[7] and Deme[8]). According to the results of these studies, soprano singers indeed tend to tune the F_1 to match the raised high f_0, if otherwise the F_1 would be exceeded by the f_0. In other words, sopranos clearly and systematically modify the production of vowels as they raise the pitch, with a modification that can be grasped as gradual opening in the acoustic vowel quality.

The harmonics of the voice source are all integer multiples of the f_0. Accordingly, as the f_0 increases, the spectrum also gets more sparse. As a result, the formant frequencies (defined as the resonances of the vocal tract, according to, eg, Stevens and House[9]) are well represented in the spectrum of the vowels in normal- or low-frequency speech, but at high f_0, the formants may not necessarily enhance the sparsely placed harmonics, and thus remain "unseen" or undetectable in the output signal. In other words, as the f_0 increases, the acoustic information carried by the spectrum decreases and the resolution of the vowels in the speech signal diminishes (see Figure 1). This effect is often referred to as the spectral undersampling effect (see eg, de Cheveignè and Kawahara[10]). The spectral undersampling effect poses a problem both to human perception and to acoustic analysis. As far as human perception is concerned, this effect implies that vowel perception must be to some extent dependent on f_0 (due to the reducing spectral information at high f_0s), but at the same time, the spectral undersampling makes it also relatively difficult to explain how vowel perception processes work at higher pitches.[10] As for acoustic analysis, due to the undersampling effect,

Accepted for publication July 14, 2016.
From the Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary.
Address correspondence and reprint requests to Andrea Deme, Research Institute for Linguistics, Hungarian Academy of Sciences, H-1068 Budapest, Hungary. E-mail: deme.andrea@nytud.mta.hu
Journal of Voice, Vol. 31, No. 2, pp. 252.e1–252.e14
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.07.008

formant estimation becomes more and more unreliable from the output signal as the f_0 increases (see Figure 1, and eg, Epps et al[11] and Wolfe et al[12]).

The combined impact of the $F_1 : f_0$ tuning and the undersampling effect results in phonetic underspecification of sung vowels at higher f_0 values. Because of this underspecification, decreased intelligibility of high-pitched sung vowels is also often supposed. In general, studies testing the intelligibility of high-pitched sung vowels provide a support for this supposition (Scotto di Carlo and Germain,[13] Benolken and Swanson,[14] Hollien et al,[15] and Deme[16,17]). However, the perceptual tendencies described in these studies are in apparent contradiction regarding the types of mismatches characteristic to the "errors" of perception. Authors of earlier studies claim that in accordance with the gradual opening observable in the production of sung vowels, these vowels are also perceived as more open as the f_0 is increased (see, eg, Scotto di Carlo and Germain,[13] Benolken and Swanson,[14] and Hollien et al[15]). If we take a closer look at these studies, however, it becomes evident that the authors draw their above summarized conclusion either (i) in spite of the fact that they find a great number of evidence contradicting it, or (ii) purely on the basis of results for close vowels (which may only be perceived as more open in the case of a stimulus-response mismatch that concerns the degree of openness). (Please note that in the present study, the vowels' degree of openness is referred to as the first, the second, the third, and the fourth degrees of openness, where the first degree corresponds to close vowels, adapted from Traunmüller.[18])

Recent studies building their experiments on more balanced material (with respect to the tested vowels' degree of openness) showed that Hungarian vowels of the second and the third degrees of openness (ie, those that could be "misidentified" in both direction on the continuum of openness) were perceived as more close in high-pitched singing below the f_0 value of about 700 Hz.[16] Based on the authors' results and conclusions, it may be implied that high-pitched singing is an ideal domain to investigate the nonlinear relations characterizing the relationships between the acoustics and the perception of speech sounds.

In previous studies, vowel perception in high-pitched singing was also investigated from the perspective of phonetic cues that may promote the identification of vowels. (It is important to note here that studies on sung vowel perception do not in all cases acknowledge the fact that the acoustic quality of the vowels change with pitch raising, and thus in these studies, stimulus-response mismatches are evaluated as "errors" in vowel identification. In the present study, however, identification is estimated in the percentage of the identification of the intended—but potentially changed—vowel quality to emphasize that acoustic vowel qualities are doubtlessly modified as the f_0 increases.)

Smith and Scott[19] tested whether high-pitched sung vowels uttered (and presented) in consonantal (CVC) context may be perceived at a higher percentage according to the intended quality than vowels uttered in isolation. For this research question, the classic literature on speech perception provided a basis. According to the literature, due to coarticulation and the resulting dynamic formant transitions, the identification of vowels is more efficient if vowels are uttered (and presented to the listener) in a C-C context (ie, in CVC sequences) than if vowels are uttered and presented in isolation.[20,21] To test the promoting effect of formant transitions in high-pitched singing, Smith and Scott[19] compared vowel identification in two conditions: (i) in minimal sets (ie, in sense words) and (ii) in "isolation" (but note that in the cited study, Smith and Scott actually cut out the steady-state portions of vowels from the CVC sequences to test the identification of isolated vowels). The results showed that C context yielded much more "identifiable" vowels than those of the steady-state portions of vowels. Based on these findings, the authors concluded that due to coarticulatory effects, sung vowels may preserve their original quality in perception even at high f_0, if they are uttered in C context. The authors did not reflect on the salient contradiction between earlier findings on acoustics and their findings on perception. They also did not elaborate an answer to the question of how listeners might be able to compensate for the acoustic quality changes accompanying pitch raising by means of their perceptual processes. Moreover, the conditions used to test the effect of the formant transitions further complicate the interpretation of the results. In the case of the CVC context, the task of the listeners was to choose the vowel that was uttered in the /b-d/ skeleton, ie, (i) to identify an existing word, and (ii) to choose one from the possible four answers *bid*, *bead*, *bed*, and *bad*. In the isolated vowel condition, however, the number of vowels as possible responses was apparently not limited to four by any restriction. Moreover, formulating the task in another way, we may also say that in the isolated vowel condition, listeners had to identify vowels in nonsense words (as opposed to the CVC condition in which sense words were to

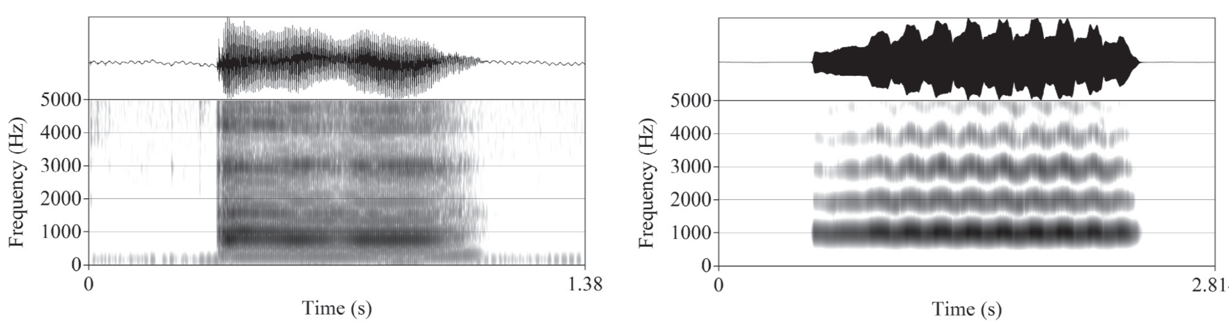

FIGURE 1. The Hungarian /aː/ in speech at approximately 200 Hz (**A**) and in singing at 988 Hz (B5) (**B**) produced by a professional soprano. At 988 Hz, the undersampling effect is observable resulting in no clear track of formants in the spectrogram.