Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks

Daniel Friedrichs[a)]
Department of Speech, Hearing and Phonetic Sciences, UCL
2 Wakefield Street, London WC1N 1PF, United Kingdom

Dieter Maurer
Institute of the Performing Arts and Film, Zurich University of the Arts (ZHdK)
Toni-Areal, Pfingstweidstrasse 96, CH-8031 Zurich, Switzerland

Stuart Rosen
Department of Speech, Hearing and Phonetic Sciences, UCL
2 Wakefield Street, London WC1N 1PF, United Kingdom

Volker Dellwo
Phonetics Group, Department of Computational Linguistics, University of Zurich
Andreasstrasse 15, CH-8050 Zurich, Switzerland

4 Figures
11 Tables (Appendix)

[a)] Author to whom correspondence should be addressed: daniel.friedrichs@ucl.ac.uk

Abstract

The phonological function of vowels can be maintained at fundamental frequencies (f_o) up to 880 Hz [Friedrichs et al. (**2015**). J. Acoust. Soc. Am. **138** , EL36–EL42]. Here, we test the influence of talker variability and multiple response options on vowel recognition at high f_os. The stimuli (n=264) consisted of eight isolated vowels (/i y e ø ɛ a o u/) produced by three female native German talkers at eleven f_os within a range of 220–1046 Hz. In a closed-set identification task, 21 listeners were presented excised 700-ms vowel nuclei with quasi-flat f_o contours and resonance trajectories. The results show that listeners can identify the point vowels /i a u/ at f_os up to almost 1 kHz, with a significant decrease for the vowels /y ɛ/ and a drop to chance level for the vowels /e ø o/ towards the upper f_os. Auditory excitation patterns reveal highly differentiable representations for /i a u/ that can be used as landmarks for vowel category perception at high f_os. These results suggest that theories of vowel perception based on overall spectral shape will provide a fuller account of vowel perception than those based solely on formant frequency patterns.

PACS number: 43.71.-k

I. INTRODUCTION

Patterns of formant frequencies are commonly assumed to be the most salient cues to vowel perception. The assumption that the vowel identification process is mainly driven by such an underlying acoustic representation contributes largely to the pervasive idea that listeners' ability to recognize vowels has to be poor at very high fundamental frequencies (f_o) due to a sparse sampling of the vocal tract transfer function. This holds true, in particular, when the normal range of the first formant frequency (F_1) is exceeded by f_o, and the higher formants are poorly specified due to a wide spacing of the harmonics.

Support for this view is mainly provided by studies on Western operatic singing. Howie and Delattre (1962), for example, found in a study on the perception of high-pitched vowels (f_o range 132–1056 Hz) sung by a baritone and a soprano that vowels lose their identity increasingly with increasing f_o. This degradation starts with the categories usually characterized by a low F_1 (i.e., high vowels such as /i/ and /u/) and leaving only those with the highest F_1 (i.e., low vowels such as /a/ and /ɑ/) identifiable at very high f_os. Ever since, numerous studies have reported that only /a/-like vowels can remain identifiable at the highest musical notes near 1 kHz (see Sundberg, 2013, p. 87, for an overview). It seems plausible, however, that this loss of vowel contrast is primarily due to articulatory changes applied by Western operatic singers when they perform at higher pitches. In experimental studies such as Joliveau et al. (2004) it has been shown, for example, that sopranos shift the first resonant frequency (f_{R1}) of their vocal tract – and thus F_1 – to the vicinity of f_o as soon as f_o drastically exceeds the normal range of f_{R1} of an intended vowel. This tuning of f_{R1} is achieved by increasing the jaw opening and reducing the maximum constriction of the vocal tract (Sundberg, 1975; Sundberg, 2013). As f_o gains considerable amplitude when being closer to a resonant frequency, these maneuvers may help a singer to maintain vocal power and timbral homogeneity (Smith and Wolfe, 2009). However, the acoustic modifications associated with shifting a resonant frequency may lead to ambiguous formant frequency patterns and consequently to a confusion of vowel categories.

Given this situation, it is surprising that few studies have investigated vowel recognition outside Western operatic singing at very high f_os as there is evidence that even a sparsely sampled vocal tract transfer function still carries information, which can be used by listeners to recognize different vowels, despite a likely absence of the supposed F_1 and an undersampling of the higher formants. Smith and Scott (1980), for example, reported listeners' identification performance significantly above chance level (mean of 70% correct) for the four front vowels /i ɪ ɛ æ/, which were produced by a soprano in isolation at an f_o of about 880 Hz (i.e., the musical note A5) with a raised larynx (i.e., a shortened vocal tract), and thus not in an articulation mode typical for Western operatic singers. When asked to produce the same vowels in her operatic singing style, identification dropped to a mean of 4% correct at the same f_o. Maurer and Landis (1996) showed that infant and adult talkers can produce identifiable versions of the vowels /i a o u/ but not of /e/ at an f_o between about 500–870 Hz that was individually chosen by the talker. In a more recent study, Maurer et al. (2014) investigated the high-pitched vowels /i y œ a ɔ u/ produced by a female Cantonese opera singer in isolation and monosyllabic consonant-vowel utterances and found that /i a ɔ u/ could be identified by more than 80% of the listeners within an f_o range of 820–860 Hz. In a study using a two-alternative forced choice task, Friedrichs et al. (2015a) provided evidence that the phonological function of the eight vowels /i y e ø ɛ a o u/ (i.e., the function they fulfil in linguistic contrastive position to help listeners distinguish between words) can be maintained at f_os up to at least 880 Hz when they were produced in minimal pairs. These judgments were made on excised steady-state vowel nuclei (250 ms) excluding consonantal context phenomena such as co-articulation and formant transitions. This is particularly surprising for vowels that typically have a low F_1 that were tested in combination with adjacent vowels with similar F_2 (e.g., /i/ vs. /e/ and /u/ vs. /o/), because an absent F_1 has been argued to make vowels with a similar F_2 indistinguishable (Smith and Wolfe, 2009, p. E196; see Ito et al., 2001, for contradictory results). In a follow-up study (Friedrichs et al., 2015b), a female talker produced the same vowels except /u/ in the German word context /l–V–gən/ (/u/ was excluded as it would have resulted in a meaningless utterance), and a multiple-choice identification task was used. It was found that the words including /i y a o/ remained identifiable – and thus the vowels' phonological function could be maintained – throughout the investigated f_o range from 220 to 880 Hz. For the vowels /e ø ɛ/, however, a significant decrease was observed in listeners' identification performance within this range (for /ø/ from about 587 Hz and for /e ɛ/ from about 784 Hz). At the highest f_o used (880 Hz), listeners could recognize the vowel /ɛ/ again.

The acoustic features and perceptual mechanisms underlying accurate vowel category perception at such high f_os remain unclear. As some of these studies found high identification rates even when excluding cues that play an important secondary role in vowel perception (e.g., vowel duration and formant frequency movement, see Lehiste and Peterson, 1961), it seems possible that spectral information apart from formant frequencies allowed listeners to identify vowels at very high f_os. Besides vowel identification models that are based on formant frequency distribution, speech scientists (in particular, from the automatic speech recognition community) have long recognized that overall spectral shape as reflected by, for example, Mel Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980), are a more robust feature set than formants. Pols et al. (1969) and Klein et al. (1970) showed that a simple filter bank analysis (essentially an auditory excitation pattern approach which encodes the overall shape of the spectrum) matched perceptual vowel spaces well. Zahorian and Jagharghi (1993) found in an automatic vowel classification experiment that spectral-shape features (the discrete cosine transform coefficients of a bark frequency scaled spectrum) are superior acoustic cues for vowel identity classification compared to formants. Ito et al. (2001) showed that also the amplitude ratio of high- to low-frequency components (i.e., the spectral tilt) affects the perceived vowel category and is at least equally effective as F_2 as a cue for vowel identification. Several overall-spectral-shape models have been advocated over the last decades (see Kiefte et al., 2013, for a more comprehensive review of this approach). Most of them do not pay special attention to the distribution of formants, but are based on the assumption that the gross shape of a smoothed spectral envelope underlies the identification process. As it is very unlikely to find common formant frequency patterns at f_os of about 880 Hz, it seems possible that the overall spectral shape – despite a severe undersampling of the spectral envelope (see de Cheveigné and Kawahara, 1999, and Hillenbrand and Houde, 2003, for more details on this problem) – might have conveyed the information that allowed listeners to identify different vowel categories (but see Maurer, 2016, for an argument that perceived vowel categories are more a result of a complex systematic interaction between spectral shapes and f_o than has generally been assumed in phonetic theory).

However, it is also possible that the lack of between-talker acoustic vowel variation facilitated identification of the vowels (excepting Maurer and Landis, 1996, who used vowels of infant and adult talkers, all of the above-mentioned studies showing accurate vowel category perception at high f_os were single-talker studies). In that situation, listeners may have adapted to the talker's individual articulatory behavior (i.e., the within-talker acoustic vowel variation). Thus, it is not clear whether the results can be generalized to other talkers and whether an experimental design including more than one talker would lead to similar results. In addition, it seems likely that the number of response options (i.e., binary and multiple-choice tasks were used) had an effect on the identification performance as listeners perform better when fewer response options are provided.

The present study addresses these issues. Here, we asked three female talkers to produce the eight vowels /i y e ø ɛ a o u/ in isolation (thus eliminating possible confounding effects due to co-articulation with adjacent consonants) at eleven f_os within a range of 220–1046 Hz. In a multiple-choice task (mixed-talker condition) with all possible vowels as response options, listeners had to identify single 700-ms nuclei with quasi steady-state acoustic characteristics. These center portions of the vowels were used to exclude possible secondary cues, in particular, sweeping harmonics in the on- and off-sets, which might sample the vocal tract transfer function more continuously and thus provide information about the position of the formants.

To investigate possible spectral properties underlying listeners' identification process at high f_os, we calculated simple versions of the excitation patterns that these vowels would be expected to generate in the auditory periphery and discuss them with respect to the results of the identification test.

II. METHODS

A. Subjects

21 native German listeners (10 female, 11 male; mean age = 23.2, s.d. = 2.25) participated in a multiple-choice vowel identification task. All were students at the University of Zurich and none of them reported any hearing impairments when asked before the experiment.

B. Stimuli and apparatus

Three female native German talkers with professional voice training (one soprano, age: 33; one Musical-Theatre singer, age: 34; one actress, age: 34) were recorded with a cardioid condenser microphone (Sennheiser MKH 40 P48 with pop shield, Wedemark-Wennebostel, Germany) on a PC via an audio interface (RME Fireface UCX, RME, Halmhausen, Germany) in a noise-controlled room at Zurich University of the Arts (ZHdK) (Switzerland). The sampling frequency of the recordings was 44.1 kHz. Subjects were recorded keeping a constant distance of about 30 cm to the microphone when standing on a drawn position reference on the floor. They were selected based on samples from a corpus of recordings of 60 talkers because of their extended vocal range and noticeable skill of maintaining vowel categories at high f_os. As part of the standard procedure as implemented in an associated project (see Maurer et al., 2016, for more details), the latter was assessed in a listening test using a blocked-talker condition and a multiple-choice identification task carried out by five phonetically trained listeners. The other 57 talkers (both female and male) had more limited vocal ranges and were not capable of producing vowels throughout the designated f_o range from 220 to 1046 Hz.

The three subjects were then asked to produce the eight long vowels /i y e ø ɛ a o u/ in isolation at eleven f_os (220, 330, 440, 523, 587, 659, 698, 784, 880, 988, 1046 Hz) with a monotone pitch contour resulting in 264 recordings (11 frequencies * 8 vowels * 3 talkers). Piano notes were presented as reference sounds to the subjects via loudspeaker immediately preceding the production. The talkers were asked to focus on producing recognizable vowels and to ignore typical voice aesthetics that might be important in their respective artistic style. The lowest f_o (220 Hz) corresponds to the female average f_o in citation-form words (Hillenbrand et al., 1995). The highest f_o (1046 Hz) corresponds to the high C (the musical note C6) in soprano singing and exceeds the normal range of F_1 of all German vowels produced by female talkers (see Pätzold and Simpson, 1997). The average f_o of each vowel was measured in Praat (Boersma and Weenink, 2016) using it's autocorrelation method (Boersma, 1993) and later checked manually. All vowels used in this study were recorded several times to ensure that at least one had an actual f_o close to the target f_o and a minimum duration of 1 second. All vowels that met these criteria were then evaluated again in the same listening test carried out by the five phonetically trained listeners, and the vowels with the highest identification scores were selected as stimuli. The mean duration of the final recordings was 1.49 s (range from on- to offset of voicing: 1.18 – 2.83 s).

Only vowel centers of 700 ms (± 350 ms from the vowel midpoint) with quasi-flat f_o contours and steady-state spectral characteristics were used as stimuli. On- and offsets of the excised sounds were faded over 5 ms by amplitude modulating the waveform with raised cosines. All stimuli were normalized to an arbitrary intensity. The overall output level was chosen by listeners individually to be comfortable.

C. Procedure

A mixed-talker listening test was carried out in a small and noise-controlled room at the University of Zurich (Switzerland) using closed dynamic headphones (Beyerdynamic DT 770 Pro, 250 Ω). The experiment consisted of a multiple-choice identification task with all 8 vowels as response options. Listeners (n=21) were presented the excised 700-ms vowel nuclei while they saw a screen that contained eight circularly arranged buttons, each button labeled with one category (randomly arranged). Above the response buttons listeners could read the question *Welchen Vokal hörst Du?* (*Which vowel do you hear?*). The listener's task was to identify the vowel presented from the eight response options provided. After listeners made their choice they heard the next stimulus automatically with a delay of one second. Listeners could not repeat a stimulus. Each listener heard each token only once which means that any particular vowel at each f_o was responded to 63 times.

D. Data analysis

We performed a set of statistical analyses on correct/incorrect responses using mixed-effects logistic regression models in R (version 3.3.1; R Development Core Team, 2016, lmerTest package; Kuznetsova et al., 2014), in which listeners and items were entered as random variables (Baayen et al., 2008). The predictors were vowel category, f_o, talker, and all their interaction. The significance of the main effects and interactions was assessed with likelihood ratio tests that compared the model with the main effect or interaction to a model without it. For clarity's sake, the results and figures are presented in percentages, although all statistical analyses were performed on raw data (correct/incorrect responses). The estimates (β) that are reported in the results section are expressed in logit units and were computed taking "incorrect response" as the reference level for the dependent variable.

To investigate possible shifts towards other than the intended vowel categories, 11 confusion matrices (one for each f_o, each based on a total of 504 samples, i.e., 8 vowels x 3 talkers x 21 listeners' responses) with the two dimensions *intended vowel* (actual class) and *response vowel* (predicted class) were calculated.

E. Excitation patterns

Simple auditory excitation patterns were generated for each vowel using a 200-channel linear gammatone filter bank, whose bandwidths and centre frequencies were calculated according to the ERB formulae given by Glasberg and Moore (1990). The rms level of the output wave was calculated for each filter channel, and converted to dB. In addition, a frequency weighting was applied to account for the transmission properties of the middle ear, as based on measurements made by Puria et al. (1997).

III. RESULTS

Results obtained from the logistic regression revealed a highly significant effect of f_o ($\chi^2(10) = 30.8$, $p < .001$), a highly significant effect of vowel category ($\chi^2(7) = 28.21$, $p < .001$), no main effect of talker ($\chi^2(2) = 2.24$, $p = .33$), and a highly significant interaction between the three ($\chi^2(244) = 627.91$, $p < .001$). For the ease of interpretation, and as a complex three-way interaction makes it impossible to ignore any one of them in accounting for the effects of the other two, we decided to break down the data into three sets to test for a two-way interaction between vowel category and f_o for the individual talkers. The results of the three analyses showed consistently a highly significant interaction between vowel category and f_o (talker 1: $\chi^2(70) = 188.42$, $p < .001$; talker 2: $\chi^2(70) = 182.74$, $p < .001$; talker 3: $\chi^2(70) = 209.5$, $p < .001$). Significant effects of vowel category were found for all talkers (talker 1: $\chi^2(7) = 28.19$, $p < .001$; talker 2: $\chi^2(7) = 22.01$, $p < .01$; talker 3: $\chi^2(7) = 35.77$, $p < .001$), and f_o (talker 1: $\chi^2(10) = 30.79$, $p < .001$; talker 2: $\chi^2(10) =$

32.61, $p < .001$; talker 3: $\chi^2(10) = 30.2$, $p < .001$). Taken together, these effects suggest that listeners' identification performance showed high variability between vowel categories and across f_os generally.

Figure 1 shows the distribution of the percentage of correct identification for each f_o and talker across vowels. Throughout the f_o range the overall performance declined more or less continuously for all talkers.

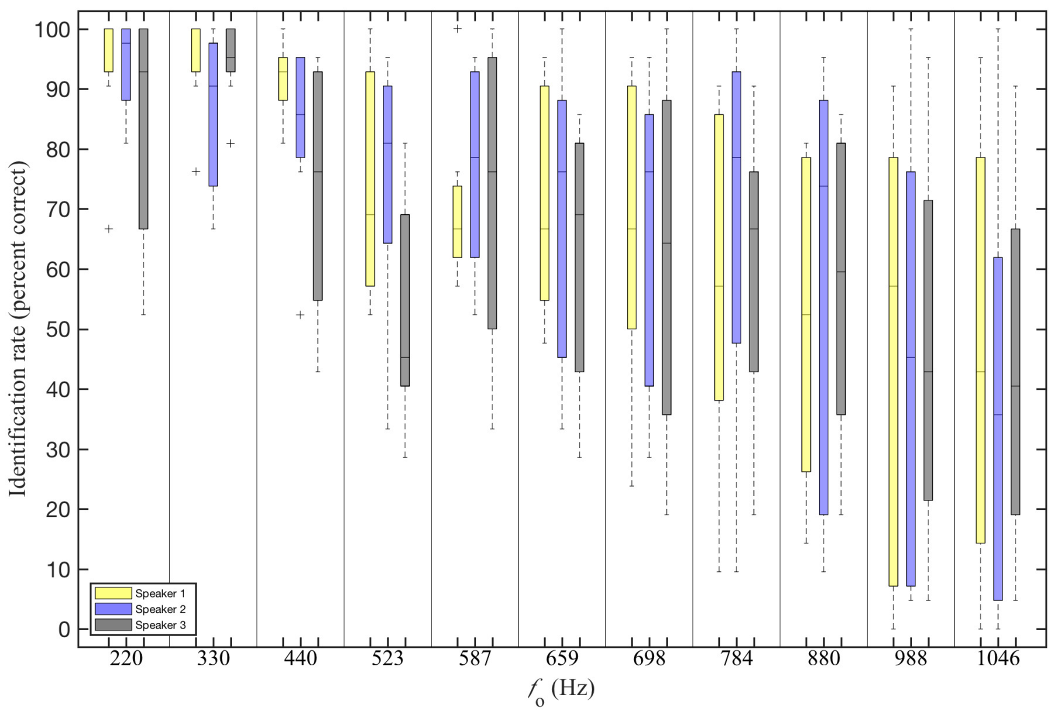

Figure 1: (Color online) Box plots showing the distribution of percent correct for the identification of all investigated vowels at the eleven f_os for the individual talkers.

The increasing variability toward the higher f_os can be explained by an increasing inter-vowel variability, as the identification rate of individual vowel categories differed greatly between low and high f_os. This can be seen in Figure 2 showing the mean percent correct scores for each individual vowel at the different f_os. Listeners' identification performance for the vowels /i ɛ a u/ is surprisingly stable up to at least 880 Hz, and percent correct values can typically be found in the range above 70%. At the two highest f_os (988 and 1046 Hz), the identification rate for /ɛ/ drops to intermediate ranges between 40 and 50% correct. Only the point vowels /i a u/ remain in the upper third of the percent correct scale. On the contrary, for the vowels /e ø o/ an extensive decrease in listeners' identification performance can be found throughout the f_os from 220 to 1046 Hz. While identification scores range between 90–100% at the two lowest f_os (220 and 330 Hz), they drop fairly continuously toward chance level for these three vowels, which is reached at 988 Hz. The identification rate of /y/ drops substantially at an f_o of 523 Hz (from about 85 to 60% correct) and decreases despite some variability towards upper f_os. From 988 Hz identification scores are similar to those of /ɛ/ (i.e., within the 35–50% correct range).

Confusion matrices (see Figure 3, for a graphical illustration; the raw data can be found in Appendix A) reveal dominant shifts toward the vowel categories /i a u/ in cases of false identifications at the highest f_os. For /ɛ/, strong confusions at the highest two f_os (988 and 1046 Hz) were found with /a/, which also showed the highest response proportions of all vowels at these f_os (28% and 24.4%). The drop in identification

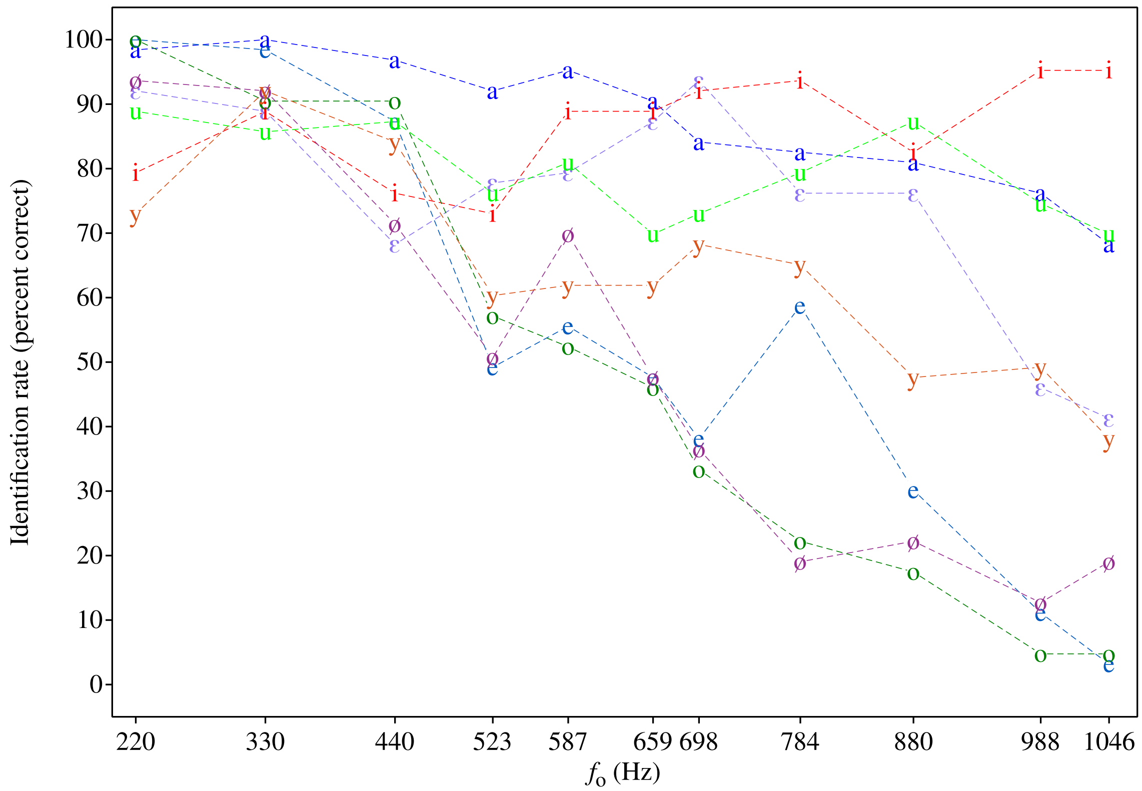

Figure 2: (Color online) Line graphs showing percent correct values, summed over all talkers, for the identification of each of the eight vowels over the investigated f_o range.

performance for the vowel /y/ in the range from 523 Hz on upwards is due to a confusion with other front vowels and from 784 Hz upwards mainly due to a confusion with /i/. A confusion between these two vowels also explains the relatively poor performance for /i/ at the lowest f_o 220 Hz (15.9% of the listeners responded /i/ when /y/ was presented to them). In case of /ø/, shifts in perception were generally found to be widely spread, that is, toward all the investigated vowel categories except /i/. The majority of false identification of /o/ shifted from a perceived /a/ at 523 and 587 Hz to /u/ at all higher f_os. Within the range 523–784 Hz, the vowel /e/ was often confused with /i/. At higher f_os the perceived vowel category shifted toward /ɛ/ and /a/.

Figure 4 shows the auditory excitation patterns for the eight vowels used in this study produced at an f_o of about 988 Hz. Both the patterns calculated for individual talkers and those averaged across talkers reveal that the point vowels /i a u/ show maximally distinct spectral shapes, which can be easily distinguished by the overall excitation level in the higher frequency region above about 1.5 kHz. The obtained confusions of the vowel categories /y e ø ɛ o/ at this f_o show a high degree of correspondence to the excitation patterns of the respective point vowels they were confused with most often. For example, the pattern calculated for /o/ shows high similarity with the pattern of the point vowel /u/, that is, a relatively low excitation level in the high frequency region. The excitation pattern of /y/ exhibits a relatively high excitation level in the high frequency region, which is also the case for the point vowel /i/. The patterns of the vowels /e ø ɛ/ show intermediate levels of excitation in the high frequency region, which is also the case for /a/, the vowel which was most often responded by the listeners when these vowels were presented to them at 988 Hz.

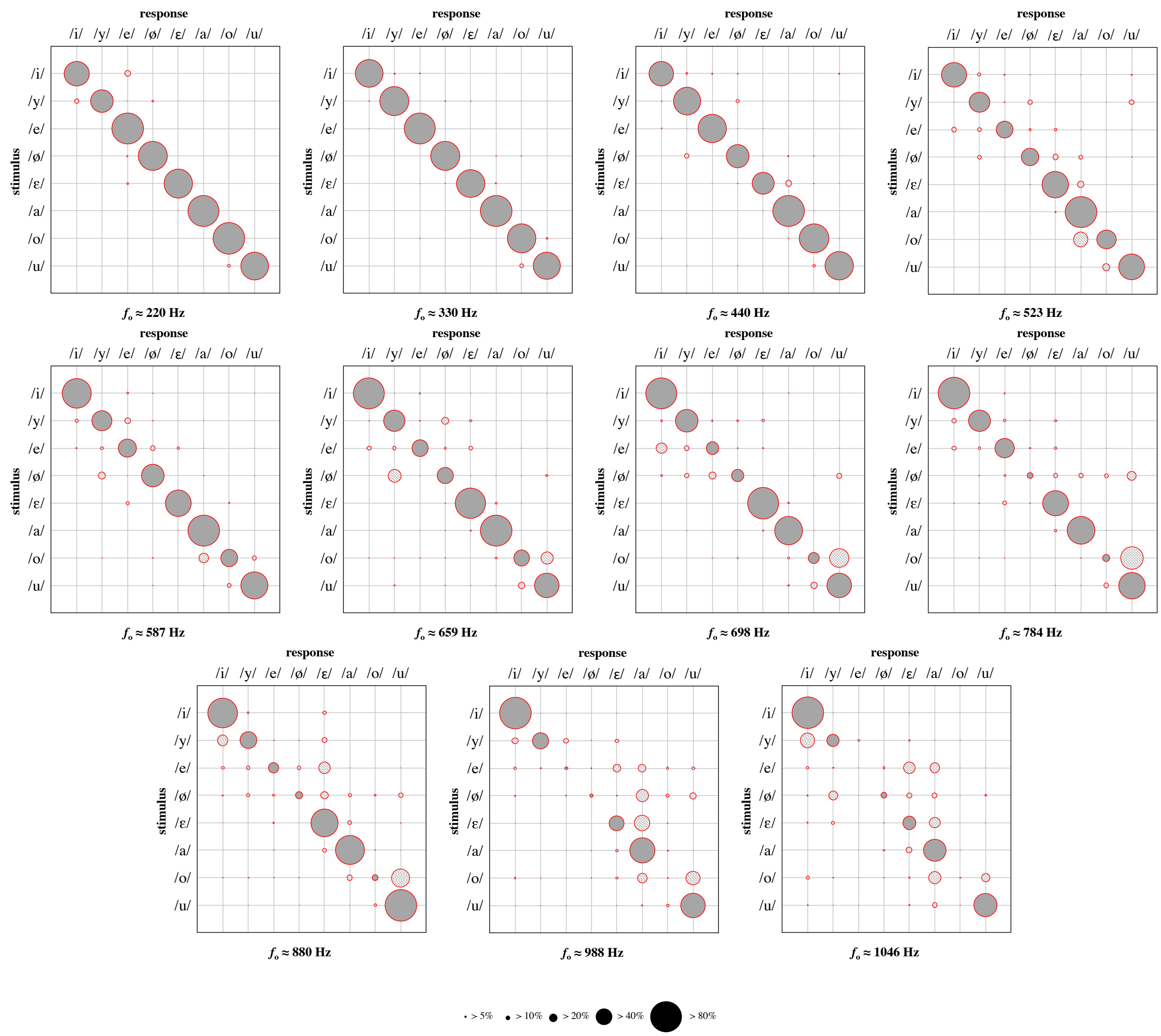

Figure 3: (Color online) Graphical confusion matrices showing the intended and response vowel categories for each f_o. The radius of each circle is proportional to the number of times that a particular stimulus (given by the row) was identified as the column response. Correct responses (down the diagonal) are solid gray, whereas identification errors (confusions) are indicated by diagonal lines through the circles.

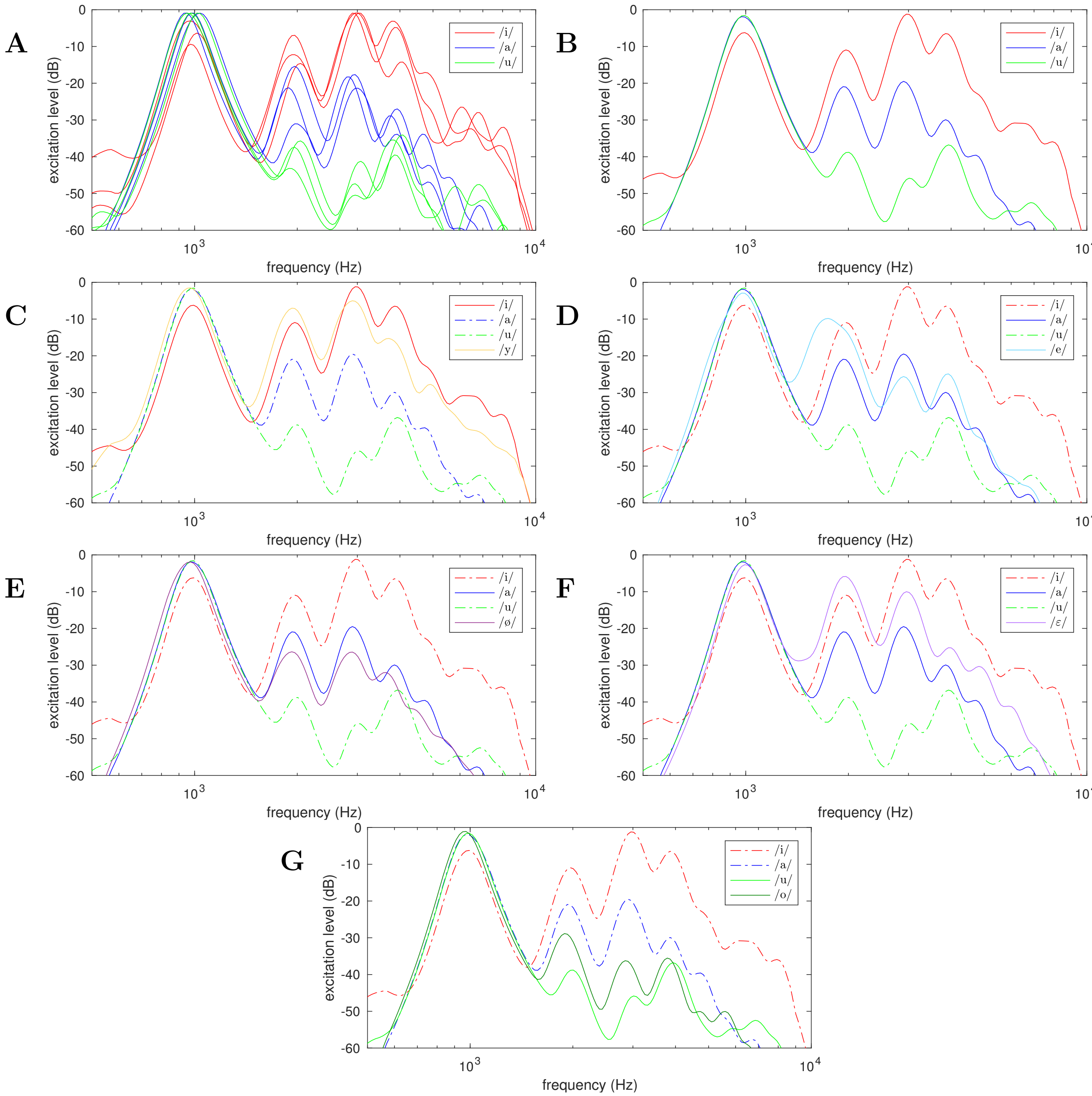

Figure 4: (Color online) Excitation patterns for the vowels used in this study that had an f_o of about 988 Hz. Part (A) shows the excitation patterns for the individual point vowels /i a u/ produced by all talkers. Part (B) shows the excitation patterns of the same vowels averaged across talkers. All other parts (C–G) show each of the other investigated vowels together with the point vowels. In these graphs, solid lines are used to indicate the strongest confusion of a respective vowel with one of the point vowels. (The information in this figure may not be properly conveyed in black and white.)

IV. DISCUSSION

The results have shown that listeners' abilities to recognize vowels within a fundamental frequency range from 220 to 1046 Hz differ greatly across vowel categories and the range of f_os. Listeners could perform well even with a variety of talkers, which means that good performance at high f_os is not being done through some odd mechanism or sensitivity which would be idiosyncratic for each talker. It is not surprising that all vowels could be identified accurately at the lowest f_os used here (220 and 330 Hz), but it is striking that only the performance for the vowels /y e ø o/, but not for /i a ɛ u/ decreased drastically within the f_o range from around 523 to 880 Hz. The results also revealed that the point vowels /i a u/ remain identifiable at an f_o close to 1 kHz or even above (in the case of /i/).

Thus, the results differ substantially from those provided by numerous studies on vowel identification in Western classical singing, which have reported consistently that high vowels such as /i/ and /u/ are the first vowels to lose their identity when f_o is progressively increased. This means that findings from the field of operatic singing cannot be generalized to other forms of speech production. In addition, the findings reported here support the hypothesis that articulatory changes which have been found in Western classical singers like resonance tuning (e.g., shifting f_{R1} to the vicinity of a higher f_o), must indeed have a strong effect on the identifiability of vowels.

Given the degree to which the vocal tract transfer function is undersampled at an f_o around 1 kHz a significant loss of formant information has to be considered as very likely (e.g., here, the vowels' typical medians of F_1 are exceeded by about 220–660 Hz, and there is only one harmonic every 1 kHz). Although it is possible that the loss of formant information can explain the decreasing identification performance, it seems likely that formants cannot be the primary acoustic correlates for vowel category perception at very high f_os.

Calculations of auditory excitation patterns for the eight vowels at an f_o of 988 Hz, revealed maximally distinct excitation levels in the frequency region above roughly 1.5 kHz for the point vowels /i a u/. Excitation patterns of the other vowels have been found to exhibit very similar spectral shapes as those of the point vowels they have been confused with most often. Both the excitation patterns of /u/ and /o/, for example, show relatively low excitation in the frequency region above 1.5 kHz, but the identification rate of /u/ (about 75% correct) was considerably higher than that of /o/ (about 10% correct), while a substantial proportion of responses (about 43%) were /u/ when /o/ was presented. As similar observations were found for other non-point and point vowel combinations, it seems likely that distinctive excitation patterns can be used by listeners as landmarks (in terms of reference points) for vowel category perception at high f_os.

Using distinctive excitation patterns as landmarks for vowel identification could also explain most of the findings reported in earlier studies on vowel identification at high f_os. Regarding the vowels used by Smith and Scott (1980) in their perception experiment (i.e., /i ɪ ɛ æ/), it is possible that the information conveyed by the distinct spectral shapes might have been sufficient for the listeners to distinguish at least between the two pairs /i ɪ/ and /ɛ æ/. However, it is difficult to draw conclusions from this as vowel duration differed substantially in this study, and not enough detail about performance with the different vowels and the instructions given to the listeners were provided.

Comparing the results of the present study to those reported by Friedrichs et al. (2015b), the diverging identification performance for the vowel /o/ is surprising. While a perfect identification rate (100% correct) was found at an f_o of 880 Hz by Friedrichs et al. (2015b), a performance near chance (17.5% correct) was observed in the present study. Although the lack of between-talker acoustic vowel variation (as being a single talker study) and secondary cues to vowel identity (vowels were presented in word context) in the former study might have helped listeners to perform better it seems possible that this difference is also due to the importance of perceptual and acoustic landmarks. The strongest support for this hypothesis is the fact that the vowel /u/ was not included in the study of Friedrichs et al. (2015b), and thus, a confusion of /o/ and /u/ like the one found in the present study was not possible (e.g., /u/ received more than 50% of the responses for the intended vowel /o/ at an f_o of 880 Hz). It seems, therefore, likely that listeners used the vowel /o/ as a substitute because /u/ was not presented to them as a response option. The results by Friedrichs et al. (2015a), who found the same eight vowels used in the present study identifiable up to an f_o of 880 Hz when recorded in minimal pairs and tested in a two-alternative forced choice task, could also be explained within this context. As a single talker was asked to produce several different two-word combinations containing a vowel in contrastive position (e.g., the German words *Buden* vs. *Boden*), it is possible that the talker produced vowels with acoustic features alike or different from those of a point vowel at higher f_os to make them distinguishable (e.g., producing an /o/ more toward /a/ to distinguish it from /u/). This way the phonological function of vowels in linguistic contrastive positions could be maintained for all vowels even at very high f_os. Given this, it is plausible that the number of response options has a strong effect on listeners' identification performance, and obviously, a better performance should be expected when fewer responses options are provided.

It is possible that the results presented here may have been driven in part by the relative frequency of German vowels. For example, in German, /i/ is more frequent than /y/, and /u/ is more frequent than /o/ (Pätzold and Simpson, 1997). Forced to choose between two vowels that otherwise match the spectral characteristics of the stimulus equally well, listeners are most likely to pick the one with the higher a priori probability. However, it is unlikely that this can explain listeners' identification performance entirely as, for example, the long /e/ is more frequent than the long /a/, with which it has been confused most often in this study at an f_o of 988 Hz. In addition, relative frequency may be the driving force behind which vowel label is applied to a cluster of similar vowels, but it cannot explain the fact that vowels were categorized into three distinct groups.

In summary, the results presented here make it clear that a theory of vowel perception based solely on formant peak patterns cannot account for the relatively preserved performance listeners demonstrate in identifying vowels at high f_os. Formal modelling of the relationship between the perceptual and physical spaces of vowels at high and low f_os are required for a convincing demonstration, but it seems likely that overall spectral shape features will play an important role in a coherent account of vowel perception generally.

Acknowledgements

This study was supported by the Forschungskredit of the University of Zurich, Grant No. FK-14-062, and the Swiss National Science Foundation (SNSF), Grants No. P2ZHP1_168375 and 100016_143943/1. Thanks to Nick Clark, whose software was used to perform the gammatone filtering, and Sandra Schwab for her helpful contributions and comments on an earlier draft of this paper.

REFERENCES

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). ”Mixed-effects modeling with crossed random effects for subjects and items,” J. Mem. Lang. **59**(4), 390–412.

Boersma, P., and Weenink, D. (**2016**). ”Praat: Doing phonetics by computer [Computer program],” Version 6.0.15, retrieved March 23, 2016 from http://www.praat.org/ (Last viewed April 30, 2016).

Boersma, P. (**1993**). “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” Proceedings of the Institute of Phonetic Sciences (17), University of Amsterdam, 97–110.

de Cheveigné, and Kawahara, H. (**1999**). ”Missing-data model of vowel identification,” J. Acoust. Soc. Am. **105**, 3497–3508.

Davis, S., and Mermelstein, P. (**1980**). Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. Proceedings of the IEEE Interiational Conference on Acoustics, Speech and Signal Processing **28**, 357–366.

Friedrichs, D., Maurer, D., and Dellwo, V. (**2015a**). ”The phonological function of vowels is maintained at fundamental frequencies up to 880Hz," J. Acoust. Soc. Am. **138** , EL36–EL42.

Friedrichs, D., Maurer, D., Suter, H., and Dellwo, V. (**2015b**). "Vowel identification at high fundamental frequencies in minimal pairs," Proc. 18th Int. Congr. Phonetic Sci., paper number 0438, 1–5.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hearing Research **47**, 103–138.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hillenbrand, J. M., and Houde, R. A. (**2003**). "A narrow band pattern-matching model of vowel perception," J. Acoust. Soc. Am. **113**, 1044–1055.

Howie, J., and Delattre, P. (**1962**). "An experimental study ofthe effect of pitch on the intelligibility of vowels," Natl. Assoc. Teachers Singing Bull. **18**(4), 6–9.

Ito, M., Tsuchida, J., and Yano, M. (**2001**). "On the effectiveness of whole spectral shape for vowel perception," J. Acoust. Soc. Am. **110**(2), 1141–1149.

Joliveau, E., Smith, J., and Wolfe, J. (**2004**). "Vocal tract resonances in singing: the soprano voice," J. Acoust. Soc. Am. **116**, 2434–2439.

Kiefte, M., Neary, T. M., and Assmann, P. F. (**2013**). "Vowel Perception in Normal Speakers," Handbook of Vowels and Vowel Disorderes, edited by M. J. Ball, and F. E. Gibbon (Taylor & Francis, LLC, New York), pp. 161–185.

Klein, W., Plomp, R., and Pols, L. C. (**1970**). "Vowel spectra, vowel spaces, and vowel identification," J. Acoust. Soc. Am. **48**(4B), 999–1009.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2014**). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0–20. Retrieved from http://CRAN.R-project.org/package=lmerTest. (Accessed on June 30, 2016)

Lehiste, I., and Peterson, G. E. (**1961**). "Transitions, glides, and diphtongs," J. Acoust. Soc. Am. **33**, 268–277.

Maurer, D., and Landis, T. (**1996**). "Intelligibility and spectral differences in high-pitched vowels," Folia Phoniatr. Logop. **48**, 1–10.

Maurer, D., Mok, P., Friedrichs, D., and Dellwo, V. (**2014**). "Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer," Proceedings of the Fifteenth Annu. Conf. Int. Speech Commun. Assoc. Singapore, 2132–2133.

Maurer, D., Suter, H., Friedrichs, D., and Dellwo, V. (**2016**). "Acoustic characteristics of voice in music and straight theatre: topics, conceptions, questions," Trends in Phonetics and Phonology. Studies from German speaking Europe, edited by A. Leemann, M. J. Kolly, S. Schmid, and V. Dellwo (Peter Lang, Bern/Frankfurt), pp. 256–265.

Maurer, D. (**2016**). *Acoustics of the Vowel - Preliminaries* (Peter Lang AG, International Academic Publishers, Bern).

Pätzold, M., and Simpson, A. (**1997**). "Acoustic analysis of German vowels in the Kiel Corpus of read speech," Arbeitsberichte des Instituts für Phonetik und Digit. Sprachverarbeitung Univ. Kiel **32**, 215–247.

Pols, L. C., Van der Kamp, L. T., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am. **46**(2B), 458–467.

Puria, S., Peake, W. T., and Rosowski, J. J. (**1997**) Sound-pressure measurements in the cochlear vestibule of human cadaver ears, Journal Of the Acoustical Society Of America **101**, 2754–2770.

R Core Team. (**2016**). R: A Language and Environment for Statistical Computing. Version 3.1.3. [Computer software] Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org (Accessed on June 30, 2016)

Smith, J., and Wolfe, J. (**2009**). "Vowel-pitch matching in Wagner's operas: Implications for intelligibility and ease of singing," J. Acoust. Soc. Am. **125**, EL196–EL201.

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

Sundberg, J. (**1975**). "Formant technique in a professional female singer," Acustica **32**, 89–96.

Sundberg, J. (**2013**). "Perception of singing," in Psychology of Music, 3rd ed., edited by D. Deutsch (Academic Press, London), pp. 69–106.

Zahorian, S., and Jagharghi, A. (**1993**). "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am. **94**, 1966–82.

Appendices

A. Confusion matrices for each f_o containing the raw data of the identification test in percentages.

$f_o \approx 220$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	79.4	0	20.6	0	0	0	0	0
/y/	15.9	73	3.2	7.90	0	0	0	0
/e/	0	0	100	0	0	0	0	0
/ø/	0	0	6.3	93.7	0	0	0	0
/ɛ/	0	0	7.9	0	92.1	0	0	0
/a/	0	0	0	0	1.6	98.4	0	0
/o/	0	0	0	0	0	0	100	0
/u/	0	0	0	0	0	0	11.1	88.9
response proportions	11.9	9.10	17.3	12.7	11.7	12.3	13.9	11.1

$f_o \approx 330$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	88.9	6.3	4.8	0	0	0	0	0
/y/	4.8	92.1	0	1.6	1.6	0	0	0
/e/	1.6	0	98.4	0	0	0	0	0
/ø/	0	0	0	92.1	0	4.8	3.2	0
/ɛ/	0	0	3.2	1.6	88.9	6.3	0	0
/a/	0	0	0	0	0	100	0	0
/o/	0	0	1.6	0	0	0	90.5	7.9
/u/	0	0	0	0	0	0	14.3	85.7
response proportions	11.9	12.3	13.5	11.9	11.3	13.9	13.5	11.7

$f_o \approx 440$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	76.2	7.9	6.3	4.8	0	0	0	4.8
/y/	4.8	84.1	0	11.1	0	0	0	0
/e/	4.8	1.6	87.3	3.2	3.2	0	0	0
/ø/	0	15.9	0	71.4	3.2	6.3	3.2	0
/ɛ/	0	0	1.6	4.8	68.3	20.6	3.2	1.6
/a/	0	0	0	0	1.6	96.8	1.6	0
/o/	1.6	0	0	0	0	4.8	90.5	3.2
/u/	0	1.6	0	1.6	0	0	9.5	87.3
response proportions	10.9	13.9	11.9	12.1	9.5	16.1	13.5	12.1

$f_o \approx 523$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	73	11.1	6.3	1.6	0	1.6	0	6.3
/y/	1.6	60.3	4.8	15.9	0	0	1.6	15.9
/e/	15.9	12.7	49.2	7.9	9.5	3.2	0	1.6
/ø/	0	12.7	1.6	50.8	17.5	12.7	1.6	3.2
/ɛ/	0	0	0	1.6	77.8	20.6	0	0
/a/	0	0	0	0	4.8	92.1	3.2	0
/o/	0	0	0	0	0	42.9	57.1	0
/u/	0	0	0	0	0	1.6	22.2	76.2
response proportions	11.3	12.1	7.7	9.7	13.7	21.8	10.7	12.9

$f_o \approx 587$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	88.9	0	7.9	1.6	0	0	0	1.6
/y/	12.7	61.9	19	4.8	0	1.6	0	0
/e/	6.3	11.1	55.6	15.9	7.9	1.6	0	1.6
/ø/	0	22.2	1.6	69.8	0	4.8	0	1.6
/ɛ/	0	0	11.1	0	79.4	0	6.3	3.2
/a/	0	0	0	0	1.6	95.2	3.2	0
/o/	0	1.6	0	1.6	0	30.2	52.4	14.3
/u/	0	0	0	1.6	0	3.2	14.3	81
response proportions	13.5	12.1	11.9	11.9	11.1	17.1	9.5	12.9

$f_o \approx 659$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	88.9	1.6	4.8	0	3.2	0	0	1.6
/y/	3.2	61.9	4.8	20.6	7.9	0	0	1.6
/e/	14.3	11.1	47.6	7.9	14.3	0	3.2	1.6
/ø/	0	38.1	1.6	47.6	1.6	1.6	1.6	7.9
/ɛ/	0	0	0	3.2	87.3	7.9	1.6	0
/a/	0	0	1.6	1.6	6.3	90.5	0	0
/o/	1.6	3.2	3.2	3.2	0	6.3	46	36.5
/u/	0	4.8	0	1.6	1.6	1.6	20.6	69.8
response proportions	13.5	15.1	8	10.7	15.3	13.5	9.1	14.9

$f_o \approx 698$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	92.1	0	3.2	0	1.6	3.2	0	0
/y/	6.3	68.3	6.3	7.9	9.5	0	0	1.6
/e/	33.3	15.9	38.1	4.8	6.3	0	0	1.6
/ø/	7.9	14.3	22.2	36.5	0	0	1.6	17.5
/ɛ/	0	0	0	0	93.7	6.3	0	0
/a/	0	1.6	3.2	3.2	6.3	84.1	1.6	0
/o/	0	0	1.6	1.6	0	6.3	33.3	57.1
/u/	0	0	0	0	0	6.3	20.6	73
response proportions	17.5	12.5	9.3	6.8	14.7	13.3	7.1	18.9

$f_o \approx 784$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	93.7	0	4.8	0	1.6	0	0	0
/y/	15.9	65.1	9.5	1.6	7.9	0	0	0
/e/	14.3	9.5	58.7	6.3	9.5	0	1.6	0
/ø/	0	3.2	7.9	19	14.3	14.3	12.7	28.6
/ɛ/	4.8	3.2	12.7	3.2	76.2	0	0	0
/a/	0	1.6	1.6	0	9.5	82.5	3.2	1.6
/o/	0	3.2	1.6	0	0	4.8	22.2	68.3
/u/	0	0	0	0	1.6	3.2	15.9	79.4
response proportions	16.1	10.7	12.1	3.8	15.1	13.1	7	22.2

$f_o \approx 880$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	82.5	6.3	0	0	11.1	0	0	0
/y/	30.2	47.6	3.2	3.2	15.9	0	0	0
/e/	9.5	11.1	30.2	11.1	33.3	3.2	0	1.6
/ø/	4.8	11.1	7.9	22.2	22.2	11.1	6.3	14.3
/ɛ/	1.6	0	6.3	0	76.2	12.7	0	3.2
/a/	0	0	3.2	0	11.1	81	3.2	1.6
/o/	3.2	4.8	3.2	4.8	0	15.9	17.5	50.8
/u/	0	1.6	0	1.6	0	1.6	7.9	87.3
response proportions	16.5	10.3	6.8	5.4	21.2	15.7	4.4	19.9

$f_o \approx 988$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	95.2	1.6	1.6	0	1.6	0	0	0
/y/	20.6	49.2	15.9	1.6	12.7	0	0	0
/e/	9.5	6.3	11.1	4.8	23.8	25.4	7.9	11.1
/ø/	6.3	1.6	4.8	12.7	4.8	38.1	11.1	20.6
/ɛ/	1.6	1.6	0	0	46	47.6	3.2	0
/a/	0	0	3.2	1.6	9.5	76.2	6.3	3.2
/o/	6.3	1.6	3.2	3.2	7.9	30.2	4.8	42.9
/u/	3.2	3.2	1.6	0	1.6	6.3	9.5	74.6
response proportions	17.8	8.1	5.2	3	13.5	28	5.4	19.1

$f_o \approx 1046$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	95.2	1.6	0	0	3.2	0	0	0
/y/	44.4	38.1	7.9	0	6.3	1.6	1.6	0
/e/	9.5	6.3	3.2	7.9	36.5	31.7	3.2	1.6
/ø/	6.3	28.6	1.6	19	17.5	17.5	1.6	7.9
/ɛ/	6.3	11.1	0	4.8	41.3	33.3	0	3.2
/a/	0	3.2	1.6	6.3	19	68.3	1.6	0
/o/	11.1	4.8	3.2	4.8	6.3	38.1	4.8	27
/u/	4.8	1.6	1.6	0	4.8	15.9	1.6	69.8
response proportions	22.2	11.9	2.4	5.4	16.9	25.8	1.8	13.7