Perceptual Identification and Normalization of Synthesized French Vowels from Birth to Adulthood

Lucie Ménard\textsuperscript{a}, Jean-Luc Schwartz\textsuperscript{a}, Louis-Jean Boë\textsuperscript{a}, Sonia Kandel\textsuperscript{b}, and Nathalie Vallée\textsuperscript{a}

\textsuperscript{a}ICP, Université Stendhal, INPG UMR CNRS No 5009, BP25 38040 GRENOBLE Cedex 9 FRANCE
\textsuperscript{b}LPE, Université Joseph-Fourier, UMR CNRS No 5105, BP4 38040 GRENOBLE Cedex 9 FRANCE
menard@icp.inpg.fr

Abstract
This paper aims at exploring the invariant parameters involved in auditory normalization of French vowels. A set of 490 stimuli, including the ten French vowels /\textipa{ɪ} \textipa{y} \textipa{u} \textipa{ɛ} \textipa{o} \textipa{e} \textipa{œ} \textipa{ɔ} \textipa{a}/ produced by an articulatory model simulating seven growth stages and seven fundamental frequency values, has been submitted as a perceptual test to 43 subjects. Results confirm the important effect of the tonality distance between F1 and F0 in perceived height. Regarding place of articulation, F2-F1, and F3-F2, in Bark, appear to be good predictors of the perceived front-back dimension. Roundedness is also examined and correlated to the effective second formant, involving spectral integration of higher formants within the 3.5-Bark critical distance.

1. Introduction

Intersubject variability involved by vowel production is large. It is well known that vowels produced by speakers with a short vocal tract (children and women) correspond to higher formant values than vowels produced by men ([1], [2]). Furthermore, fundamental frequency decreases during growth. Considering this important variation, traditional vowel characterization by the three first formant values faces several problems. [1] report that formant values of ten American English vowels uttered by men, women and children partially overlap, perceivers correctly identify the ten vowels. The overlap in the F1/F2 and F2/F3 acoustic spaces. Despite this overlap, perceivers correctly identify the ten vowels. The question therefore arises: what are the parameters involved in the identification of distinct phonological categories?

2. Invariance and normalization

During the past decades, several studies attempted to identify and deal with interindividual variability found in vowel production. At the perceptual level, these normalization procedures all aim at reducing intraclass variability and dispersion of some parameters by seeking invariant determinants of vowel classes. In the present study, we shall focus on invariance at the acoustic level.

Regarding height, the role of F0 in the normalization process has been confirmed by various experiments. [3] reports that the difference between F1 and F0, in Bark, is a nearly invariant correlate of perceived height, for Bavarian synthesized vowels. In [4], a binary classification of [+high] and [-high] American English vowels occurs at a value of F1-F0 corresponding to 3-3.5 Bark, known to be the critical distance of the center of gravity effect [5]. However, F1-F0 is not found to be a better predictor of the entire dimension of openness, for American front ([6]) and back vowels ([7]). Indeed, some categories would be related to F1 alone.

Higher formants, according to [4], would account for the perception of the front-back feature, for American English vowels. A good classification is achieved by this parameter, front vowels corresponding to a value of F3-F2 lower than 3 Bark and back vowels, to a F3-F2 greater than 3 Bark. However, [8], for Swedish, distinguishes between front and back vowels on the basis of F2-F1 values. The difference between F2 and F0 would be a correlate of place of articulation for [9].

Finally, proposals for models of the effective second formant (F’2) have shown that this parameter, involving the center of gravity effect, related to the phenomenon of perceptual integration of close formants, would be related to perceived vowel rounding ([10]). The role of F0 as a normalizing parameter, in the group F’2-F0 ([11]), or F2-F0 ([10]), is also suggested.

3. Method

The following experiment was designed to determine the main acoustic parameters involved in the perceptual normalization of French vowels uttered by various speakers, from birth to adulthood. To assess the relevance of attested normalizing factors in the identification of French vowels, we synthesized, with an articulatory model integrating non uniform vocal tract growth, the ten French oral vowels /\textipa{ɪ} \textipa{y} \textipa{u} \textipa{ɛ} \textipa{o} \textipa{e} \textipa{œ} \textipa{ɔ} \textipa{a}/ at different growth stages and different F0 values. Table 1 displays feature analysis of French oral vowels.

<table>
<thead>
<tr>
<th></th>
<th>Front (Unrounded)</th>
<th>Front ( Rounded)</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>i</td>
<td>y</td>
<td>u</td>
</tr>
<tr>
<td>Mid-high</td>
<td>e</td>
<td>œ</td>
<td>o</td>
</tr>
<tr>
<td>Mid-low</td>
<td>ë</td>
<td>œ</td>
<td>ë</td>
</tr>
<tr>
<td>Low</td>
<td>a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Feature analysis of French oral vowels.

3.1. Stimuli

3.1.1. Overview of the model

Stimuli consist in 5-formant vowels generated by formant synthesis with the articulatory model developed by S. Maeda (\textit{Variable Linear Articulatory Model} [12]), which integrates knowledge acquired from previous models with the growth data currently available ([13]). The growth process is introduced by modifying the longitudinal dimension of the vocal tract according to two scale factors, one for the anterior part of the vocal tract and the other for the pharynx, interpolating the zone in-between. Fundamental frequency
values evolve following the data gathered by [14]. This model is controlled by seven parameters, directly interpretable in terms of functionally organized articulatory blocks (protrusion and labial aperture; jaw height; tongue body, dorsum and tip position; larynx height). The use of such a model enabled us to generate various stimuli while carefully controlling articulatory and acoustic coherence.

3.1.2. Formant patterns

Vocal tracts representative of the following ages were simulated: 0, 2, 4, 8, 12, 16, and 21 years old. For each growth stage, articulatory-acoustic prototypes for the ten French oral vowels /i y u o o e æ œ a/ were determined using the concept of Maximal Vowel Space (hereafter MVS). If the entire input space of command parameters is explored – while satisfying the conditions for vowel production – one can simulate the maximal F1/F2/F3 acoustic space appearing at the output. All possible oral vowels are thus situated within the limits of this space. By comparing the 7 MVS generated by VLM, we situated the four focal vowels /i y u o/ (hereafter MVS).

Figure 1 displays the resulting values of the stimuli. The optimal formant triplets were determined for the 10 vowels (F1/F2 and F2/F3 planes) of the MVS, for each growth stage, using the concept of Maximal Vowel Space (hereafter MVS). If the entire input space of command parameters is explored – while satisfying the conditions for vowel production – one can simulate the maximal F1/F2/F3 acoustic space appearing at the output. All possible oral vowels are thus situated within the limits of this space. By comparing the 7 MVS generated by VLM, we situated the four focal vowels /i y u o/ (hereafter MVS).

Formant synthesizer was excited by a glottal waveform while satisfying the conditions for vowel production (F(36, 1512) = 25.26, p < 0.01). Accordingly, one can suspect the influence, on correct identification, of an acoustic parameter affected by vocal tract length (namely, the formants), in relation with F0. Mean correct identification measures analysis of variance (ANOVA) [vocal tract length (7) x F0 values (7)] was performed on the global results and revealed a significant effect of both vocal tract length (F(6, 252) = 153.39, p < 0.01) and F0 (F(6, 252) = 33.96, p < 0.01) on the percentage of correct identification.

4. Results

4.1. The Bark scale

First, frequency values, in Hertz, were converted into a Bark scale, using the conversion formula proposed by [16]: $F_{bark} = 7*\text{asinh}(F_{Hz} / 650)$. We also transformed the frequency data according to Syrdal and Gopal's modifications [4], to represent Traunmüller's corrected scale [3]. Prior to the Hertz-to-Bark conversion, frequency values were corrected as follows:

- frequency values below 150 Hz are raised to 150 Hz
- for frequency between 150 Hz and 200 Hz: $F_r = F - 0.2 (F - 150)$
- for frequency between 200 Hz and 250 Hz: $F_r = F - 0.2 (250 - F)$

were $F_r$ is the corrected frequency, in Hz, and $F$ is the original frequency, in Hz. These values will be referred to as the "low-frequency end corrected" values.

4.2. Analysis of identification scores

First, results were considered according to their correct identification. For each F0 and vocal tract length (represented by a given growth stage), the number of tokens perceived according to the intended stimuli was determined. A repeated measures analysis of variance (ANOVA) [vocal tract length (7) x F0 values (7)] was performed on the global results and revealed a significant effect of both vocal tract length (F(6, 252) = 153.39, p < 0.01) and F0 (F(6, 252) = 33.96, p < 0.01) on the percentage of correct identification. An interaction of vocal tract length and F0 was also significant (F(36, 1512) = 25.26, p < 0.01). Accordingly, one can suspect the influence, on correct identification, of an acoustic parameter affected by vocal tract length (namely, the formants), in relation with F0. Mean correct identification numbers, as a function of growth stage, for the 7 F0 values, are plotted in Figure 2. A noticeable difference in the shape of the curves is observable, with maximum scores appearing at different F0 values, for increasing ages: 450 Hz (newborn), 360 Hz (2 years old), 300 Hz (4 years old), 240 Hz (8 years old), 210 Hz (12 years old), 110 Hz (16 and 21 years old). A small mismatch appears, for growth stages from 8 years old to 16 years old, for which these "best" F0 values are slightly lower than the theoretical values (section 3.1.3).
ANOVA \[F0 (7) \times \text{vocal tract length (7)}\] revealed a main effect of perceived place of articulation: front (/i y u/), mid-high (/e o ø)/, mid-low (/æ ø æ/), and low (/u/). A repeated measures ANOVA \[F0 (7) \times \text{vocal tract length (7)}\] revealed a main effect of F0 \[F(6, 252) = 30.11; p < 0.01\], but no effect of F0. However, an effect of the interaction of these two factors appeared \[F(36, 1512) = 3.33; p < 0.01\]. A study of the correlations between the percentage of perceived front vowels and the following spectral parameters was then performed, all values in Bark: F2, F2-F0, \((F2-F0)+F1)/2, \text{F2-F1, F3-F2}. Except for \((F2-F0)+F1)/2, all parameters were highly correlated to perceived frontness, F2-F1 \((r = 0.84)\) proved to be the best predictor of perceived frontness, front vowels being associated to F2-F1 greater than 5.5 Bark, whereas back vowels correspond to F2-F1 values lower than 5.5 Bark. F3-F2 was also attributed a fairly good correlation coefficient \((r = 0.82)\). The proportion of perceived front and back vowels classified with the F2-F1 boundary of 5.5 Bark correspond respectively to 99.6% and 97.8%.

To summarize, Figure 3 displays the dominantly perceived vowels in the F1-F0 vs F2-F1 plane. The perceived height boundaries (2 and 4 Bark, dashed lines) as well as the perceived place of articulation boundary (5.5 Bark, solid line) are represented.

Owing to the low F1 associated to the adult vocal tract, for close vowels, combined with high F0 values of 450 Hz, the F1-F0 parameter sometimes results in negative values, as can be observed on Figure 3. It might seem theoretically misleading to represent as a perceptual cue the negative difference between F1 and F0, that is, to conceive that the energy peak located at F1 is identified despite the lack of harmonics. Furthermore, in that case, the distance between the two lowest harmonics is 3.4 Bark, possibly yielding the perception of two separate peaks. Other analysis were carried out to evaluate the contribution of several cues in the F1 region, which would represent the effective first formant, that is, the perceived F1 value. It appears that both the most prominent harmonic in the vicinity of F1, and the centroid of the first two harmonics, could represent the effective F1.

Regarding the rounding feature, a repeated measures ANOVA \[F0 (7) \times \text{vocal tract length (7)}\] of perceived rounding suggested a main effect of F0 \[F(6, 252) = 10.23; p < 0.001\] and vocal tract length \[F(6, 252) = 165.01; p < 0.01\], as well as a noticeable effect of the interaction of the two factors \[F(36, 1512) = 3.91; p < 0.01\]. The data were grouped based on perceived vowels for which a rounding distinction was
identified (that is, /ɪ y e o e œ /). We computed F2 for each vowel. The algorithm is described in [17].

Several parameters were studied by linear regression: F2-F1, F3-F2, F'2, and F2-F0. The best classification performance (96.7% for unrounded and 96.9% for rounded vowels), achieved by F2, is displayed on Figure 4, with the 15-Bark boundary, along the F2 dimension (dashed line).

![Figure 4. Dominantly perceived vowels, in the F1-F0 vs. F'2 plane.]

5. Discussion

Finally, we performed a linear discriminant analysis on the dominantly perceived vowels, with different parameters for the three features:
- F1, F1-F0 (for height)
- F2, F2-F0, F2-F1, F3-F2 (for place of articulation)
- F'2-F0, F'2 (for rounding)

It was observed that F1-F0, F2-F1, and F2 achieved the best classification, with a mean value of 91% tokens correctly classified. Following this, one can suggest that from the traditional 4-dimension acoustic space (F1, F2, F3 and F4), a three-dimensional "perceptual" space is represented by the following correlates, for French vowels: F1-F0, F2-F1 and F'2. The transformation to this space involves, first, a transformation of Hertz values into critical band units (or Bark), and then, a difference between acoustic parameters, for height and place of articulation. For rounding, a spectral integration mechanism occurs, based on the patterns of distance between F2, F3 and F4. The data do not support the hypothesis of a universal threshold of 3-3.5 Bark, corresponding to the critical distance of spectral integration, used to perform a binary classification of the [+high] versus [-high] vowels ([4]). Nevertheless, these perceptual invariants can be interpreted as perceptual constraints imposed on speech production. Further investigation, carried out on children's produced vowels, is in process.

6. Conclusion

This paper intended to determine invariant acoustic correlates of French vowels through synthetic vocal tract growth normalization. Perceptual analysis showed that the distance between F1 and F0 in Bark is a nearly invariant parameter of perceived vowel height. The front/back dimension is determined by F2-F1 or F3-F2, and rounding can be related to the effective second formant (F'2). These analysis are of great interest for the development of normalization procedures. Furthermore, results can shed light on the issue of perceptual invariance.

7. References