PATTERNS OF HUNGARIAN VOWEL PRODUCTION AND PERCEPTION WITH REGARD TO SUBGLOTTAL RESONANCES

Tamás Gábor Csapó[1], Tekla Etelka Gráczi[2], Zsuzsanna Bárkányi[2], András Beke[2,3] and Steven M. Lulich[4]
[1]Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
[2]Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[3]Department of Phonetics, Eötvös Loránd University
[4]Psychology Department, Washington University, Saint Louis, MO 63108
e-mail: csapot@tmit.bme.hu, graczi@nytud.hu, bzsu@budling.nytud.hu, beke.andras@gmail.com, slulich@wustl.edu

Abstract

Subglottal resonances (SGRs) have been reported to divide vowels into certain contrasting natural categories: low – non-low; front – back; front unrounded non-low – other front. This role of the subglottal resonances has been investigated for a handful of languages in the speech of adults and children. The present paper aims to consider the patterns of Hungarian vowels with regard to the SGRs in two different speaking styles (isolated nonsense words and spontaneous speech), in automatic vowel classification, and in speech perception. In one experiment, formant spaces of six speakers were analyzed to investigate the possible role of SGRs in dividing the vowel space into discrete regions corresponding to distinctive features. In a second experiment, automatic formant-based vowel classification was applied and extended with normalization based on the subglottal resonances, in order to determine whether knowledge of SGRs may improve automatic vowel classification. Finally, in a pilot experiment, the perceived backness of the vowel [ɔ], as a function of F2, and Sg2 is investigated in order to determine the role of Sg2 in perception of spontaneous speech.

1 Introduction

1.1 The phonetics-phonology interface and subglottal resonances

Although in early generative phonology (Chomsky & Halle, 1968) most distinctive features were defined based on articulatory (e.g. [+/– high]) or acoustic (e.g. [+/– diffuse]) parameters, a direct link between phonology and phonetic interpretation was generally eschewed, and until recently, this has remained the defining paradigm in phonological theory. However, some non-mainstream phonological approaches and theories have arisen – and in the past decades have gained in importance – which aim to anchor distinctive features and/or constraints

explicitly in phonetic and/or other functional factors. Such factors are acknowledged to play a direct role in shaping and explaining sound patterns and phonological processes (e.g. Liljencrants & Lindblom 1972; Stevens 1972, 1989, 1998; Stevens and Keyser 2010; Ohala 1975, 1983; Hume & Johnson 2001; Hayes et al. 2004, to mention just a few). One of the most successful theories which attempts to define the phonetic basis of phonological distinctive features and constraints is Quantal Theory (QT) (Stevens 1972, 1989, 1998; Stevens and Keyser 2010).

Quantal Theory relies on the claim that "the relation between an acoustic parameter that can be observed in the sound and an articulatory parameter that can be manipulated by a speaker takes a particular non-monotonic form" (Stevens 1998: 1). That is, in some regions of articulatory-acoustic space, small articulatory movements lead to large acoustic changes, while in other regions large movements lead to small acoustic changes. These latter regions are 'stable' and are supposed to underlie distinctive features, and phonological systems are thought to use the acoustically unstable regions as dividing lines between the + and – value of various distinctive features.

Some unstable regions (for vowels) have been claimed to arise from acoustic coupling of the vocal tract to the subglottal airways. When a formant and a subglottal resonance (SGR) have similar frequencies, the vowel spectrum around the formant can be significantly affected by the presence of the SGR, frequently resulting in additional formant-like peaks, attenuation of the formant, discontinuities in formant trajectories, or a combination of these. Since the subglottal system (the tracheobronchial tree) does not have moving articulators, the SGRs are fairly constant for a given speaker. The lowest three resonances of the subglottal system with closed glottis have been observed to be about 600, 1550, and 2200 Hz for adult males, with bandwidths that are in the range of 200 to 400 Hz (see Ishizaka et al. 1976, Cranen & Boves 1987, Stevens 1998). These frequencies are generally higher for female and child speakers, and can show individual differences from one speaker to another. The subglottal system can influence the sound output from the mouth opening most prominently when the glottis is open and unobstructed. When there is normal vocal fold vibration, acoustic coupling to the subglottal cavities increases and decreases periodically throughout the glottal cycle. Stevens (1998) showed that subglottal resonances can distort nearby spectral peaks, and this appears to be especially true when F2 is near the 2nd subglottal resonance (Sg2). Therefore, it has been hypothesized (Stevens, 1998) that speakers will avoid putting vowel formants in the region of the subglottal resonances. Stevens (1998) claims that Sg2 is therefore a natural division between front and back vowels. Although the effect of the lowest subglottal resonance (Sg1) on the vowel system is generally not as pronounced as that of Sg2 (because the partially open glottis introduces acoustic losses that reduce the degree of prominence of this lowest resonance), it still seems to play a role, namely that "the lowest subglottal resonance is at a frequency that tends to be lower than F1 for low vowels and higher than F1 for non-low vowels" (Stevens, 1998: 303), that is to say, it is a natural division line between low and non-

low vowels. Lulich (2010) found that the third subglottal resonance (Sg3) may form the boundary between tense and lax front vowels in English. Similarly, Csapó et al. (2009) showed that Sg3 separates the unrounded front non-low vowels from other front vowels in Hungarian. For the same vowel produced at different times or in different contexts, it has been claimed that formants are free to vary within the frequency bands defined by the subglottal resonances (Madsack et al. 2008). Lulich (2010) also reported a connection between SGRs and stop consonant places of articulation.

In addition to these production-oriented studies, recent work has shown that SGRs are salient in speech perception (Lulich et al., 2007) and useful in speaker normalization for automatic speech recognition (Wang et al., 2008, 2009). However, the relationship between SGRs and vowel formants has so far been examined in only a handful of languages, and for relatively small numbers of speakers. Chi & Sonderegger (2007), Lulich (2010), and Jung (2009a) all tested adult speakers of English (Lulich, (2010) and Jung (2009a) also tested child speakers of English); Madsack et al. (2008) tested adult speakers of two German dialects; Jung (2009b) tested adult speakers of Korean; and Csapó et al. (2009) tested four adult speakers of Hungarian. Results from these studies have been in general agreement, namely, that Sg1 divides low and non-low vowels, Sg2 divides front and back vowels, and Sg3 divides front tense unrounded and front lax vowels. On the other hand, some discrepancies have been noted. For instance, the vowels [ɛ] and [ɔ] appear sometimes to have F1 frequency distributions lower than Sg1, sometimes higher, and sometimes centered on Sg1. The low vowel [a] consistently has an F2 frequency lower than Sg2 in English, but either lower or higher than Sg2 in German, Korean, and Hungarian, depending on the consonant context and the speaker. In addition, some back vowels (especially [u] and [ʌ]) have F2 very close to Sg2 under some conditions. Some of these discrepancies may be due to inaccurate measurements of the SGRs, especially Sg1, which can be particularly difficult to identify in spectra of subglottal acoustics recordings. Other discrepancies may have arisen from insufficient numbers of subjects (for example, in Hungarian, only four speakers have previously been analyzed) or differences in speaking styles (such as read laboratory speech or (quasi) spontaneous speech).

In the present paper, we extend the research on SGRs in Hungarian by analyzing speech production data from six new speakers producing both isolated nonsense words and spontaneous speech. (The method of recording and analyzing subglottal acoustics during speech production is described below in Section 2.1.) We hypothesized that the SGRs would divide vowels into categories as previously found, but that this separation would be more categorical in nonsense words, since spontaneous speech tends to be characterized by a greater degree of coarticulation and reduction (Lindblom 1990). In addition to direct analysis of vowel formant and SGR frequencies, we tested the potential use of SGRs in automatically classifying Hungarian vowels. The extent to which the use of SGRs improves automatic vowel classification may be considered evidence that SGRs do divide the vowel space into

distinct regions corresponding to the distinctive features. It may also be possible to improve automatic speech recognition systems (at least under certain conditions, such as ASR with limited data, cf. Wang et al., 2009) by incorporating such SGR-based classification schemes. Finally, we performed a speech perception pilot experiment using vowels excised from the spontaneous speech of two speakers. This last experiment was meant to extend earlier results showing that Sg2 affects the perception of vowel backness in isolated nonsense words (Lulich et al., 2007).

1.2 The vowel system of Hungarian

The vowel system of Standard Hungarian contains 14 vowels (all monophthongs) that correspond to 7 pairs, phonologically differentiated by length, as shown in Table 1 (Siptár & Törkenczy, 2000: 51).

Table 1. Phonological classification of Hungarian vowels

	front				back	
	unrounded		rounded			
	short	long	short	long	short	long
high	i	iː	y	yː	u	uː
mid		eː	ø	øː	o	oː
low	ɛ				ɔ	aː

Phonetically, the picture is more varied (Figure 1). The high vowels in Table 1 may indeed differ in length without any difference in quality, but true minimal pairs are not easy to find, e.g. *int* ‘beckon’ vs. *ínt* ‘tendon.acc’. Mády & Reichel (2007) found that high vowels have partially overlapping distributions of duration, and discrimination between them is not possible on the basis of F1 and F2. Furthermore, in a perception test, speakers did not differentiate them clearly. Mid rounded vowels, on the other hand, besides the length distinction, always differ somewhat in their quality. In addition, the vowels [ɛ] vs. [eː] and [ɔ] vs. [aː] (written *e, é, a* and *á*, respectively) never contrast in length without significantly differing in quality as well. It can still be claimed that phonologically these vowels form pairs because they exhibit the same length alternations as other mid and high vowels (see Table 2).

Table 2. Examples for vowel shortening

víz—vizek	‘water—pl.’	*kő—kövek*	‘stone—pl.’
tűz—tüzek	‘fire—pl.’	*kéz—kezek*	‘hand—pl.’
kút—kutak	‘well—pl’	*nyár—nyarak*	‘summer—pl.’
ló—lovak	‘horse—pl.’		

Note also that *á* is generally realized as a low central vowel, or even as a low front vowel (Kovács, 2004; Beke & Gráczi, 2010), although phonologically it behaves as a back vowel. As shown in Table 2, it alternates with back [ɔ] (*nyár—nyarak*); and in vowel harmony, stems containing *á* take back suffixes, e.g. *ház+ban* 'house.inessive' vs. *kert+ben* 'garden.inessive'. It has previously been found that the low back vowel [a], if not contrasting with a low front vowel (e.g. American English [æ]), may be produced with F2 at higher frequencies than Sg2, depending on the speaker (Madsack et al., 2008) and the phonetic context (Jung, 2009a). For all four Hungarian speakers reported in Csapó et al. (2009), *á* consistently had F2 higher than Sg2. Therefore, we will consider *á* to be phonetically front in this paper.

Figure 1 shows the hypothesized relation of SGRs and vowels in Hungarian. As Csapó et al. (2009) showed for isolated nonsense words, Sg2 lies at the boundary between back and front vowels. Sg1 similarly divides low and non-low vowels, while front rounded and front unrounded vowels can be differentiated by Sg3.

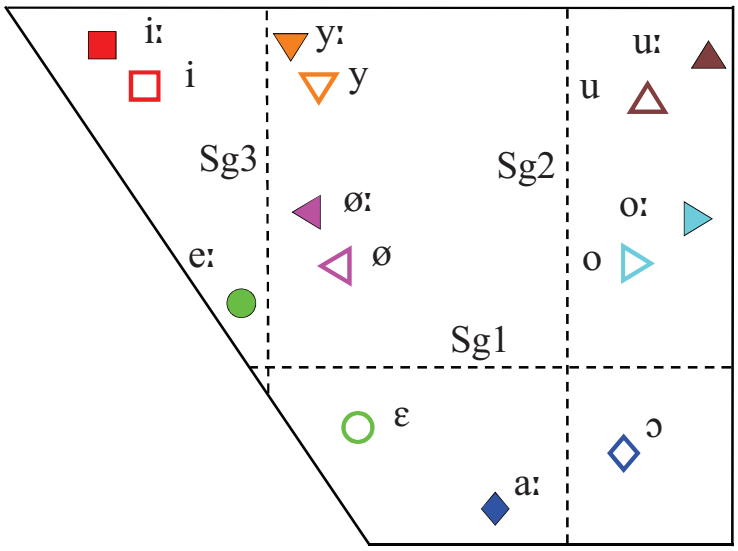


Figure 1. Hypothesized vowel space of Hungarian vowels. The horizontal line indicates Sg1, the right vertical line indicates Sg2 and the left vertical line indicates Sg3

In this paper we present data from three experiments. In section 2, we present results from two acoustic studies of Hungarian vowel formants and SGRs in both isolated nonsense words and spontaneous speech. The first study investigated the relation of vowel formants and SGRs in vowel spaces. The second study investigated the use of SGRs in the automatic classification of vowels. In section 3, we present results from a preliminary perception experiment aimed at examining the role Sg2 plays in vowel perception.

2 Acoustic experiments

In this section, we present the results of experiments on the acoustics of isolated nonsense words and spontaneous speech samples with regard to SGRs. After describing the methods used for recording and analyzing the speech samples, we present measurement results of SGRs and formants, and discuss their relation to each other. We also present results from a classification experiment in which SGR-based vowel classification is compared to standard decision tree methods.

2.1. Methods

Since previous studies of SGRs have predominantly focused on non-spontaneous speech, we made recordings of both isolated nonsense words and spontaneous speech by six adult speakers of Standard Hungarian (ages between 25 and 35 years; 5 males and 1 female). The subjects were selected from the BEA (**Be**szélt nyelvi **a**datbázis) Hungarian spoken language database. This database consists of several different speech recordings for each speaker (for further details see Gósy 2008, or http://www.nytud.hu/dbases/bea/index.html). In this experiment, we chose a quasi-monologic interview for each speaker, in which they talked about their job and/or their education. The interviewer was the same person (a 27 year-old woman) during each recording session. These recordings of the database were 2-5 minutes long. Subjects produced 543–1630 vowels, which were appropriate for formant analysis. Since in Hungarian no significant vowel reduction appears in non-accented syllables, all vowels were included in the present experiment regardless of syllable accent, except for vowels realized with creaky voice in which the formants were not well defined. All together, 5948 vowels were measured in the six subjects' spontaneous speech. The same subjects were later asked to read isolated nonsense words from a list. The nonsense words consisted of 3 syllables /ɔ/CVC/ɔ/, where the two consonants were the voiced plosives (/b, d, g/). Each plosive appeared both at the first and the second place for each target vowel (e.g. /ɔbɔbɔ/, /ɔdɔdɔ/, /ɔgɔgɔ/, /ɔdɔbɔ/, /ɔbɔgɔ/, etc). All Hungarian vowel phonemes were tested, so that each speaker read a list of 126 nonsense words once. The target vowels appeared in the second syllable. The spontaneous and nonsense word recordings were carried out in an anechoic chamber with an Audio-Technica AT 4040 microphone. The subglottal signal was recorded in a separate session with a K&K HotSpot accelerometer attached to the skin of the neck (held by the hand) below the thyroid cartilage, as described below. Both the microphone and accelerometer recordings were made with a 44 kHz sampling rate and 16 bit quantization on a computer using a Realtek sound card.

The speech samples were segmented automatically into sounds with a Hungarian forced-alignment program (Mihajlik et al. 2002). The segmentations were corrected manually in Praat based on visual and audio inspection. The F1 and F2 data of each target vowel were collected automatically from the microphone signal with a Praat script measuring at the midpoint of the vowel. The values that differed by more than 10% from the mean of the vowel group were checked and remeasured manually in Praat. To make recordings of subglottal acoustics, subjects read a short paragraph, or

summarized a paragraph which they had previously read during the recording. The SGRs were measured from the LPC of the subglottal signal with Wavesurfer (Sjölander & Beskow 2009). The measurement of SGRs is similar to reading off formants, as shown in Figure 2, since SGRs give rise to spectral peaks in the accelerometer signal, just as formants do in the microphone signal. Since SGRs are roughly constant for a given speaker, it is possible to obtain an estimate of the SGRs from a small set of measurements. Twenty measurements were made of Sg1, Sg2 and Sg3 for each speaker. The measurements were made in non-creaky sonorant sounds, approximately at their midpoints.

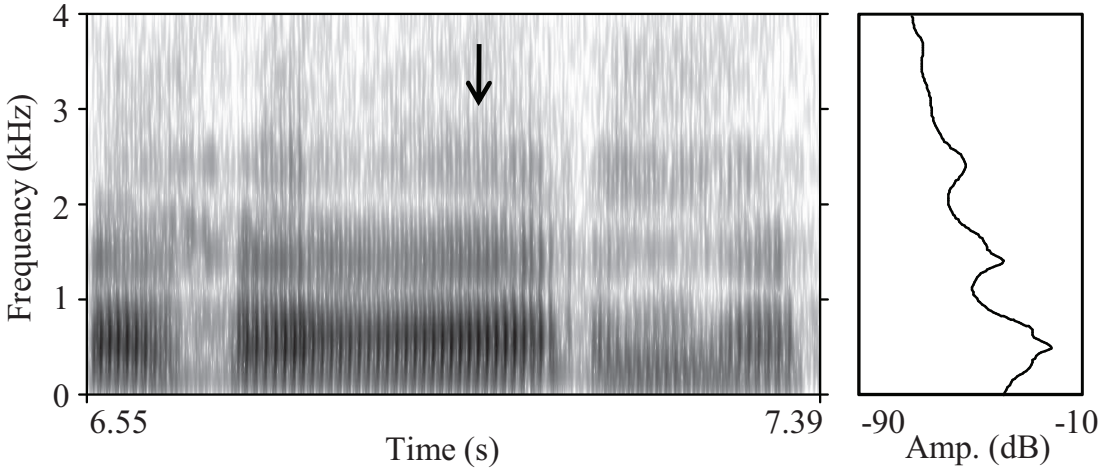


Figure 2. Sample accelerometer spectrogram and LPC spectrum from the subglottal signal of speaker M1 (resampled at 8kHz). The LPC was taken at the place indicated by the arrow. Spectral peaks of the LPC spectrum correspond to SGRs. At low frequencies (0–800 Hz), the LPC spectrum has additional peaks near Sg1, presumably due to the first few harmonics

The role of SGRs was tested in several types of analyses on vowel-realizations extracted from spontaneous speech and isolated nonsense words. Our main goals were to analyze i) how accurately the SGRs divide the formant data into the above-mentioned categories (see Figure 1); and ii) how this hypothesized function of the SGRs is applicable in automatic vowel classification.

The first goal was investigated by analyzing the vowel spaces of the six speakers separately for isolated nonsense words and spontaneous speech. Optimal separation thresholds between vowel classes were also determined using receiver operating characteristics (ROC) curves.

Automatic vowel classification

The second goal (in determining whether SGRs may help vowel classification) was tested by normalizing the formant frequencies and applying decision tree-based classification to vowels extracted from the spontaneous speech. In these experiments, either the normalized or raw formants of the vowels were the input

parameters and the classes of the previously investigated distinctive features were used as the targets. Three types of classifications were performed:

I) Formant normalization to an SGR,

II) Decision tree classification with formant normalization to an SGR,

III) Decision tree classification based on raw formant data (without normalization to an SGR).

The following phonological distinctive features were tested with the above-mentioned classifications:

a) low–non-low,

b) back–front,

c) front unrounded non-low–any other vowels.

In I), the formants of the vowels were frequency-normalized based on SGRs. For instance, in I.a), the raw F1 values for each speaker were normalized with respect to Sg1 (denoted as Fn1 = F1/Sg1) and then pooled together. After that, it was examined whether the low vowels have Fn1 > 1.0 and non-low vowels have Fn1 < 1.0. Analyses of I.b) and I.c) were done similarly: in I.b), raw F2 values were normalized with respect to Sg2 (Fn2 = F2/Sg2); in I.c), raw F2 values were normalized with respect to Sg3 (Fn3 = F2/Sg3).

In II), frequency normalization was applied as in I). After that, the open-source J4.8 binary decision tree (an improved version of the C4.5 binary decision tree;Witten & Frank 2005) was applied with 10-fold cross-validation in Weka. For instance, in II.a), the input of the decision tree was the frequency-normalized formants (Fn1), the target was the binary class of low – non-low. In this simple case, the decision tree has only one fork and two leaves. The algorithm calculated the optimal threshold of the classification (denoted by T). For the optimal threshold, T, the ratio of low vowels having Fn1 > T and non-low vowels having Fn1 < T is the highest among all possible thresholds. The results of the 10-fold cross-validation were averaged. Similarly, the input of II.b) was Fn2 with the back – front target; the input of II.c) was Fn3 having front unrounded non-low – any other vowels as target.

In III), a J4.8 decision tree was applied as in II), but without frequency normalization. The raw formant data were used as input for the classification, having the phonological distinctive feature classes as target. For instance, in III.a), the F1 values of all vowels and all speakers were pooled together and the optimal threshold T between the low – non-low vowel classes was calculated using the decision tree. 10-fold cross-validation was applied and the results were averaged. In III.b), the decision tree was used with raw F2 as input and back – front as target, while in III.c), the same was done with raw F2 as input and the above-mentioned c) classes as target.

2.2. Results and discussion

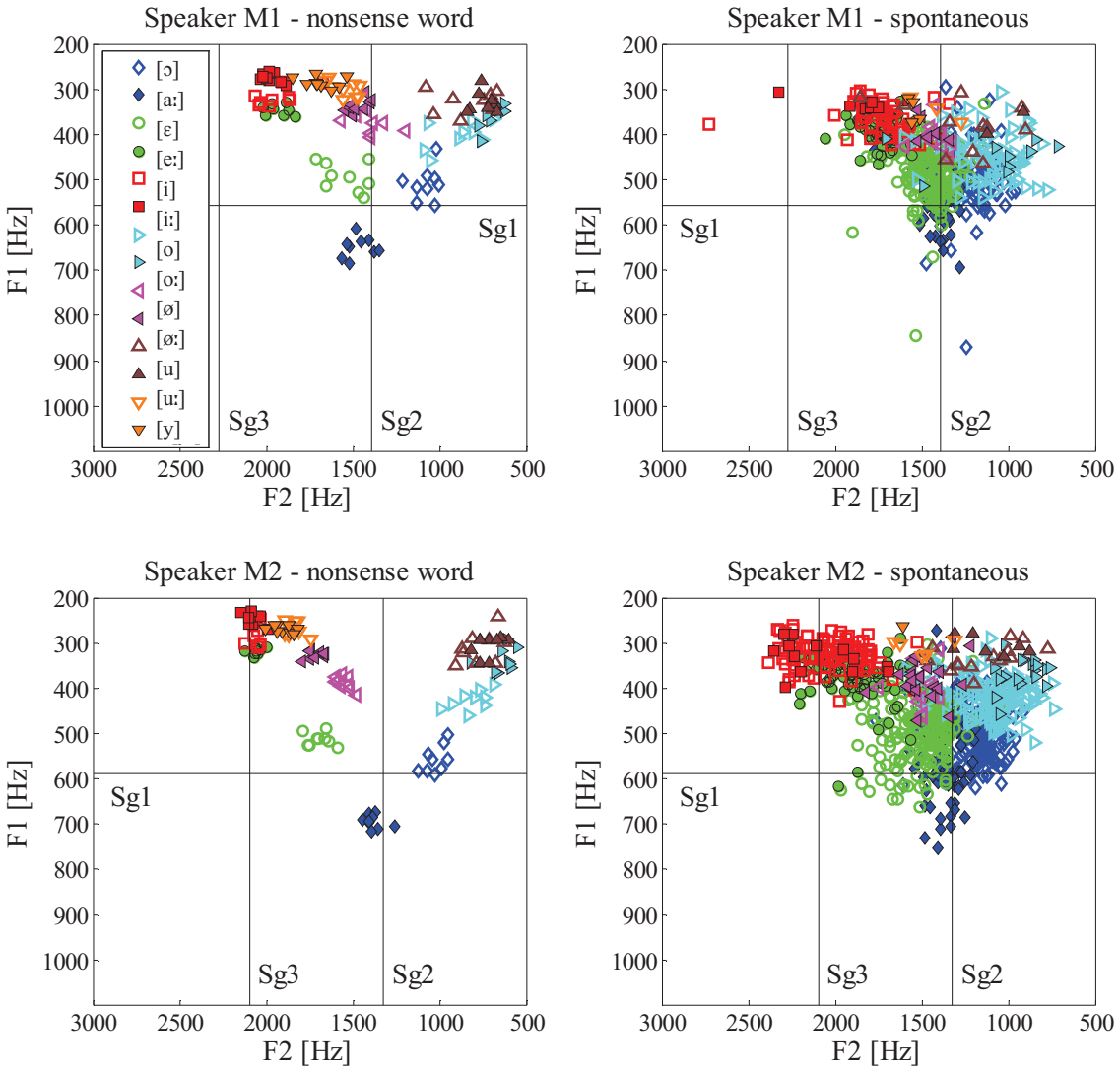
Acoustic measurements

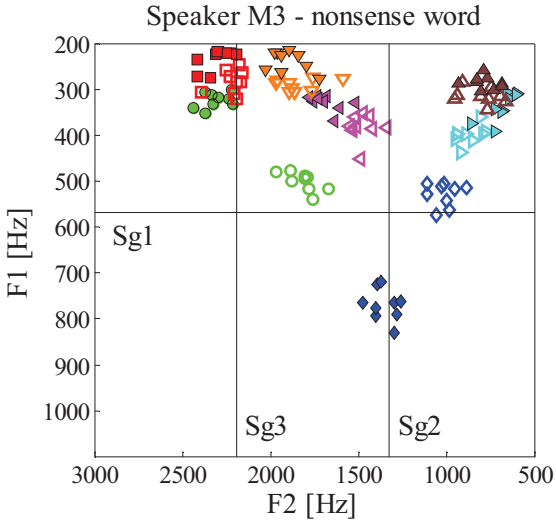
Twenty measurements of the SGRs were made for each speaker. The median values are shown in Table 3. Standard deviations were in the range of 50–100 Hz.

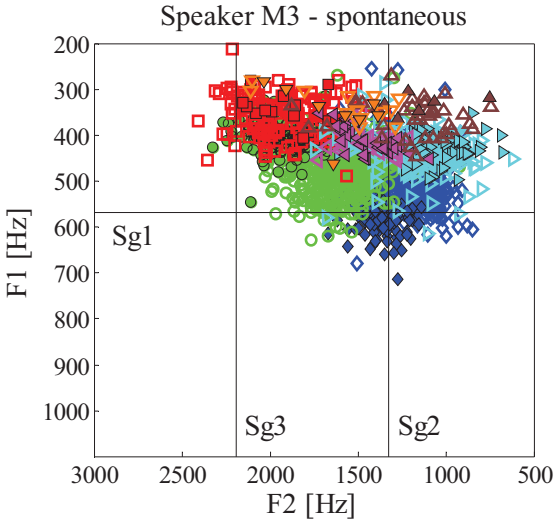
Table 3. Median values of subglottal resonances for the six speakers (all numbers in Hz)

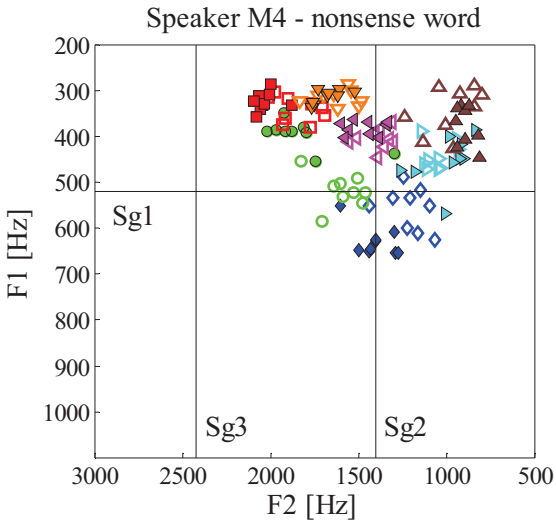
Speaker	Sg1	Sg2	Sg3
M1	556	1392	2273
M2	587	1326	2096
M3	567	1326	2192
M4	521	1402	2420
M5	545	1299	2193
F1	558	1532	2354

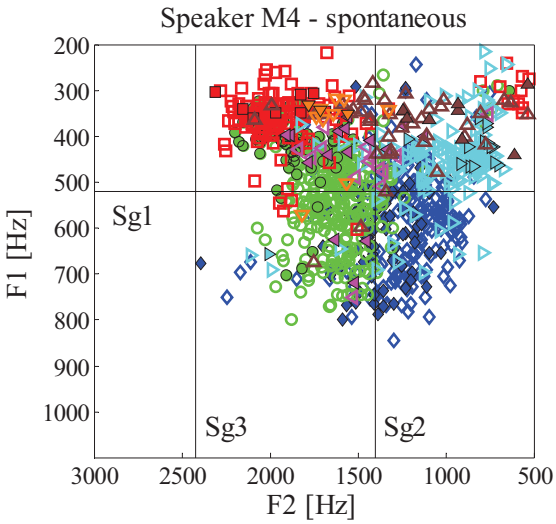
Vowel plots for each speaker are shown in Figure 3. The vowel identities are given for speaker M1, but the same legend applies in all of the panels. Straight horizontal and vertical lines correspond to the SGRs. The left column of panels shows formant spaces based on isolated nonsense words, while the right column corresponds to spontaneous speech.

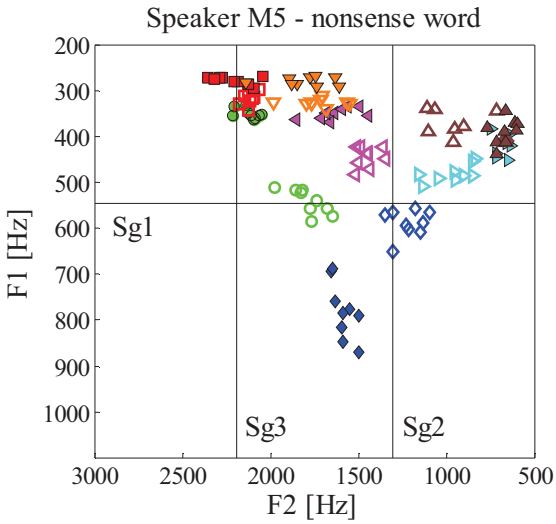


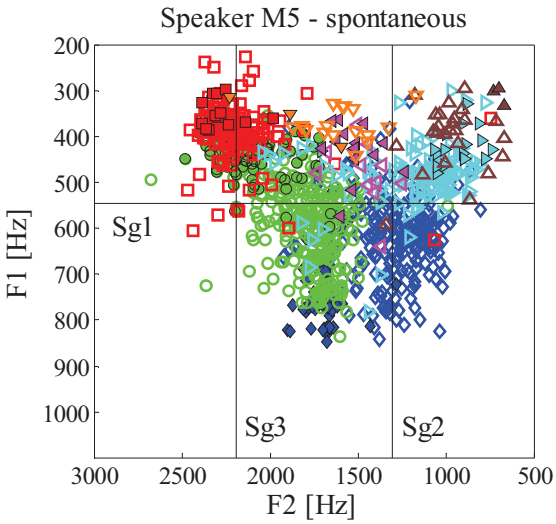

Speaker M3 - nonsense word
F1 [Hz]
F2 [Hz]
Sg1
Sg3
Sg2


Speaker M3 - spontaneous
F1 [Hz]
F2 [Hz]
Sg1
Sg3
Sg2


Speaker M4 - nonsense word
F1 [Hz]
F2 [Hz]
Sg1
Sg3
Sg2


Speaker M4 - spontaneous
F1 [Hz]
F2 [Hz]
Sg1
Sg3
Sg2


Speaker M5 - nonsense word
F1 [Hz]
F2 [Hz]
Sg1
Sg3
Sg2


Speaker M5 - spontaneous
F1 [Hz]
F2 [Hz]
Sg1
Sg3
Sg2

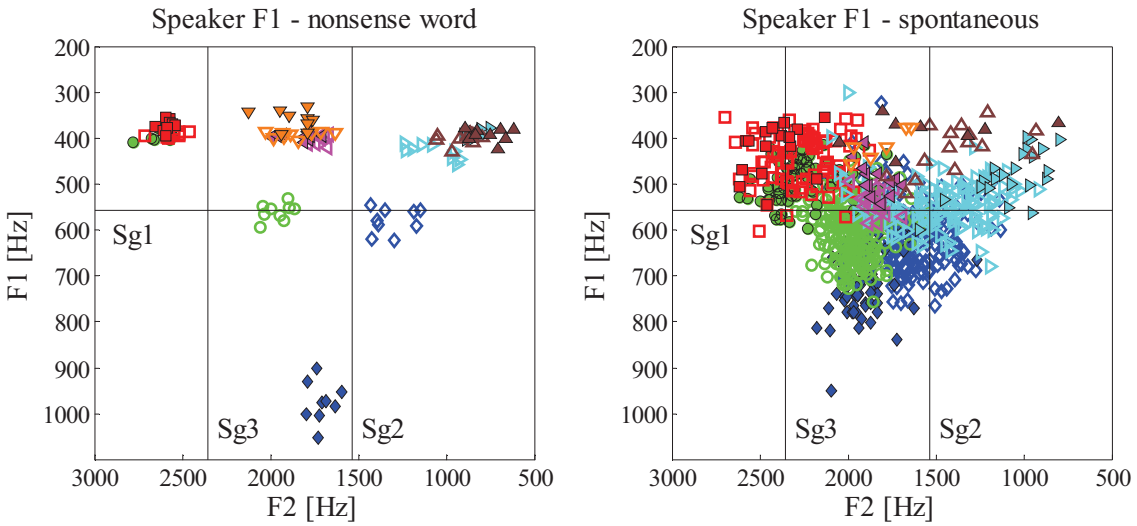
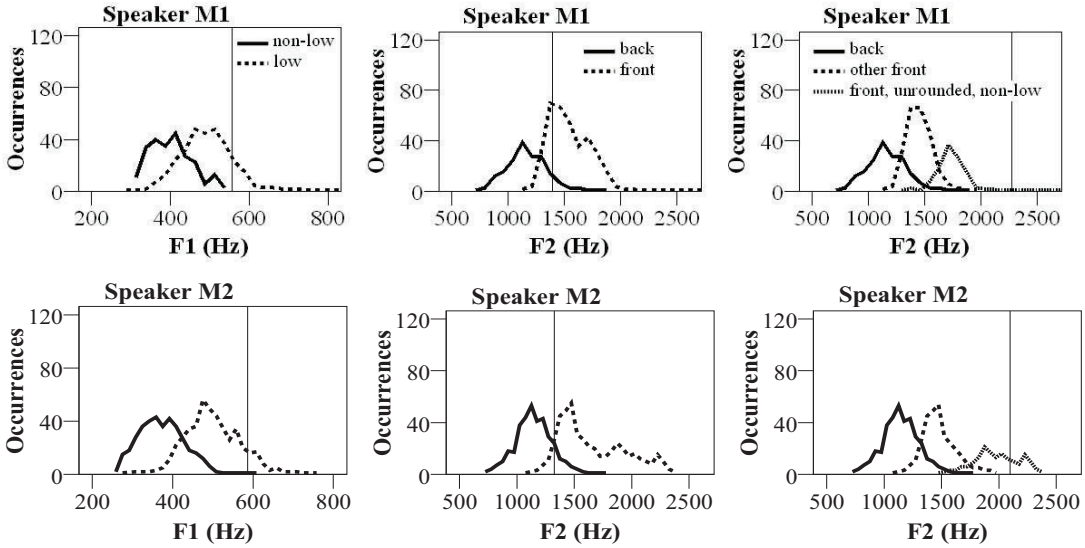


Figure 3. Vowel plots for the six speakers. Horizontal and vertical lines indicate SGR frequencies. Empty symbols represent short vowels, and filled symbols are long vowels. The left column of panels is for isolated nonsense words data; the right column is for spontaneous speech

Isolated nonsense words

In general, the separation of vowels by Sg2 in isolated nonsense words is clear, while the effects of Sg1 and Sg3 are speaker dependent. The back – front separation caused by Sg2 (right vertical line) is clearly visible. Only the short front rounded vowel [ø] and the long low vowel [aː] sometimes have an F2 lower than Sg2 for two speakers (M1 and M4). For speakers M1, M2, and M3, non-low vowels including [ɔ] and [ɛ] have F1 lower than Sg1, whereas speakers M4, M5, and F1 have [ɔ] and [ɛ] first formant frequencies near or higher than Sg1. Sg3 separates the front high unrounded vowels from the other front vowels only for speakers F1, M3, and to a small extent M5.



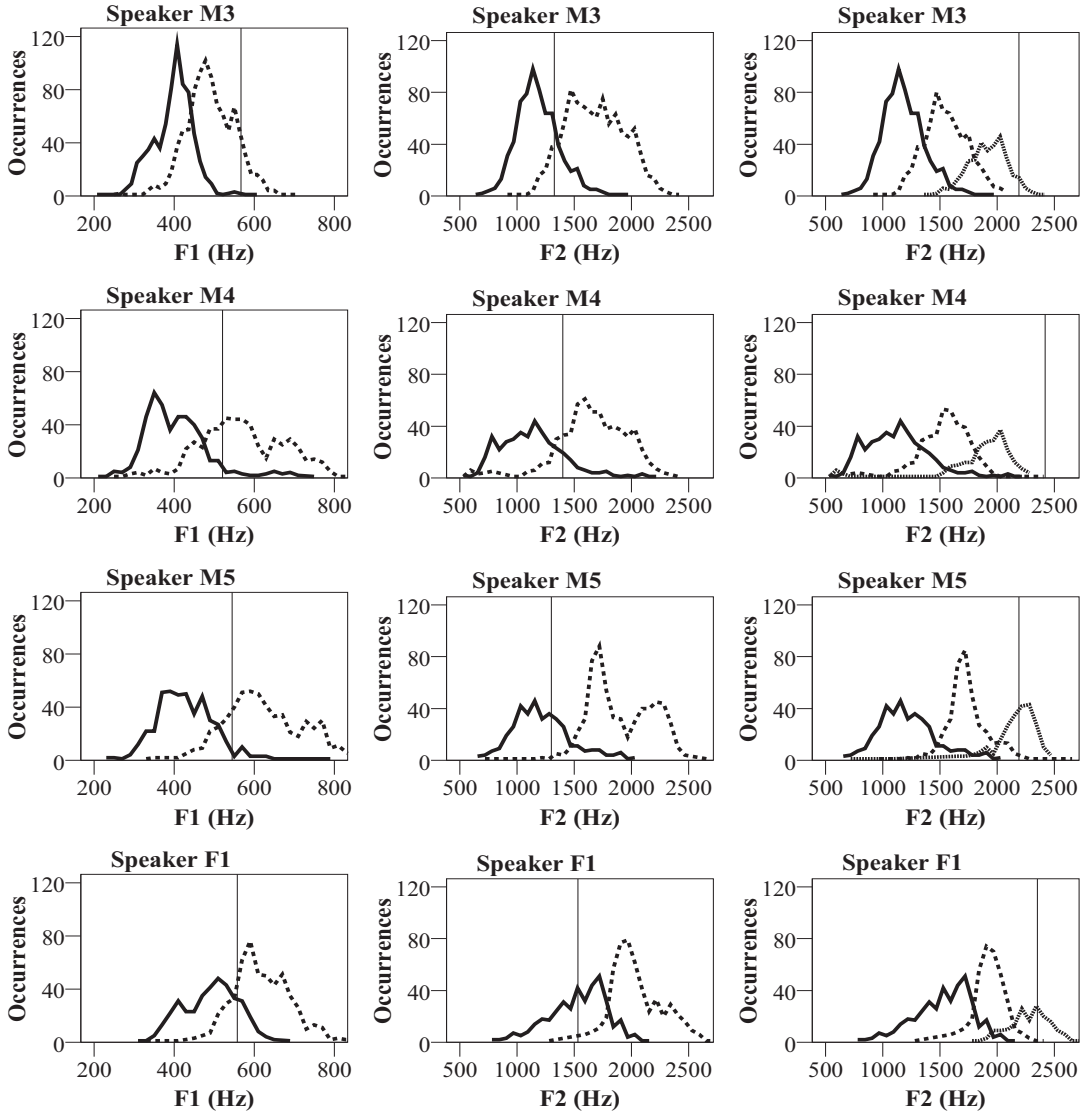


Figure 4. Vowel distribution of all speakers corresponding to the formant – SGR relation. First column: F1 – Sg1, second column: F2 – Sg2, third column: F2 – Sg3. (The vertical solid lines represent the median SGRs.)

Spontaneous speech

As expected, the results are similar but somewhat less straightforward in spontaneous speech. Figure 4 shows the distribution of the formants of the vowel categories relative to the SGRs.

In the first column, the F1-Sg1 relation is represented for the low–non-low contrast. While 95.07% of the non-low vowels had lower F1 than Sg1 across all speakers, the behavior of the low vowels was speaker dependent. As in the case of the isolated nonsense words, speakers M1, M2, and M3 had F1 lower than Sg1 in the vowels /ɛ/ and /ɔ/, whereas for speakers M4, M5, and F1, these vowels had Sg1 higher than F1. The vowel /aː/ was realized on the expected side of the Sg1 in 72.81% of all cases across speakers, but less frequently for speakers M1, M2, and

M3. Therefore the total distribution of low vowels (/ɛ/, /ɔ/, and /aː/) is speaker dependent, with the speakers falling into one of two groups.

The second column of Figure 4 represents the speaker by speaker relation of the F2 values to the Sg2 on the basis of the contrast back – front. Consistent with the nonsense word vowel spaces in Figure 3, the F2-Sg2 relation in spontaneous speech shows a more categorical division of the back – front categories across speakers than the F1-Sg1 relation does for the low – non-low categories. 74.97% of the back vowels and 91.79% of the front vowels were realized on the expected side of the Sg2. However, for speaker M5, and especially speaker F1, a large number of [ɔ] and [o] vowels occurred with F2 greater than Sg2, and for speaker M4, a number of [aː] productions had F2 lower than Sg2.

The third column of Figure 4 shows the distributions of the F2 frequencies of the vowels according to their relation to Sg3. The front unrounded non-low vowels were expected to be realized with higher F2 than the speaker's Sg3. As a general rule, this hypothesis was clearly not supported. Only the vowel /iː/ had F2 frequently higher than Sg3, and this was true of only two speakers (M2, M5, F1).

Based on these results, it appears that Sg2 divides front and back vowels most consistently. Sg1 divides low and non-low vowels, with some speaker-dependent variation in the production of the mid vowels [ɔ] and [ɛ]. As noted above, Sg1 is frequently the most difficult SGR to measure in accelerometer signals, and it is possible that inaccuracies in these measurements led to some of the variation seen in the mid vowels [ɔ] and [ɛ]. However, given the distribution of [ɔ] and [ɛ] F1 frequencies, a more likely explanation appears to be that the position of these vowels relative to Sg1 is speaker dependent. Sg3 does not consistently divide front unrounded high vowels from other front vowels. In practically all cases, the separation of vowels by SGRs is less categorical in spontaneous speech than in isolated nonsense words, but follows the same general patterns.

ROC analysis

We analyzed receiver operating characteristics curves (ROC; Cardillo, 2008) for each SGR and speaker separately, and for the two speech styles (graphs not shown). The optimal thresholds or ranges for low–non-low (Sg1), back–front (Sg2) and unrounded non-low front–other front (Sg3) categories are shown in Table 4. Those optimal thresholds that are within one standard deviation of the respective SGRs are denoted with an asterisk. Note that we assume the vowels /ɛ/ and /ɔ/ are low vowels, even though for speakers M1, M2, and M3, they have F1 categorically lower than Sg1.

The Sg1 +/– one standard deviation overlaps with the optimal range for the nonsense words and spontaneous speech of speakers F1 and M4. The optimal separation of back and front vowels in spontaneous speech falls close to Sg2 in three cases out of six. In nonsense words, a weaker trend was found. This contradicts the general assumption that the role of SGRs is clearer in laboratory speech. Sg3 shows a strong separating effect only in the nonsense word reading of two speakers (M3 and F1).

Table 4. Results of ROC analysis (all numbers in Hz)

SGR	Speaker	SGR +/– 1 SD	Nonsense word		Spontaneous	
			Optimal threshold (range)		Optimal threshold	
Sg1	M1	515–597	416		448	
Sg1	M2	548–626	460–489		424	
Sg1	M3	535–599	452–477		442	
Sg1	M4	427–615	478	*	477	*
Sg1	M5	515–575	510		509	
Sg1	F1	514–602	457–532	*	555	*
Sg2	M1	1337–1447	1213		1342	*
Sg2	M2	1261–1391	1125–1259		1343	*
Sg2	M3	1277–1375	1111–1256		1394	
Sg2	M4	1341–1463	1261		1402	*
Sg2	M5	1256–1342	1345		1569	
Sg2	F1	1490–1572	1427–1594	*	1808	
Sg3	M1	2216–2330	1765		1562	
Sg3	M2	2007–2185	1955		1620	
Sg3	M3	2119–2265	2029–2161	*	1702	
Sg3	M4	2284–2556	1677		1712	
Sg3	M5	2106–2280	1979		1893	
Sg3	F1	2247–2459	2122–2459	*	2046	

Our hypotheses regarding the connection between vowels and SGRs are therefore only partially supported by this analysis, as in the previous vowel space and formant distribution analyses. Although Sg2 appears to be closely related to the boundary between back and front vowels, Sg1 has a weaker effect on the separation of vowels: the distinction between low and non-low vowels in terms of Sg1 is speaker dependent, with the vowels [ɔ] and [ɛ] sometimes falling below Sg1 and sometimes above. Sg3 separates high front unrounded vowels from other front vowels only in two speakers and only in isolated nonsense words. Sg3 does not appear to play a general role in dividing these classes of vowels.

We predicted that the role of SGRs in dividing the vowel space would be less categorical in spontaneous speech than in isolated nonsense words. The data with regard to Sg1 have a similar pattern for both nonsense words and spontaneous speech, whereas the optimal thresholds between the back–front categories in spontaneous speech are in all cases higher than in nonsense words, and the spontaneous speech thresholds are closer to Sg2. As a consequence, the separating

role of Sg2 was found to be stronger in spontaneous speech. The role of Sg3 is, as noted above, not prominent.

SGRs and vowel classification

The results of the three types of automatic vowel classifications described in Section 2.1 will now be discussed. We investigated whether knowledge of SGRs can improve the automatic classification of vowels.

a) Low–non-low classification

Table 5. Results of the low – non-low classification

	I (SGR)		II (SGR + decision tree)		III (decision tree)	
	correctly classified	mis-classified	correctly classified	mis-classified	correctly classified	mis-classified
non-low	95.07%	4.93%	68.57%	31.43%	70.48%	29.52%
low	43.32%	56.68%	85.57%	14.43%	87.13%	12.87%
total	66.07%	33.93%	78.09%	21.91%	79.81%	20.19%
threshold	1.0		0.7966		442 Hz	

The results of the low–non-low classification are given in Table 5, showing the percentages of correctly classified vowels and misclassified vowels (the weighted average percentages are given in the row labeled “total”). The thresholds determined for the three types of classification are given in the last row.

The frequency normalized F1 with 1.0 as threshold (I) gives fairly good results for the non-low category, but misclassifies more than half of the low vowels, as was also found from visual inspection of the vowel spaces of the individual speakers. However, the decision tree classification with formant normalization (II) set the optimal threshold 20.34% lower than the Sg1, yielding an improved 78.09% accuracy. The decision tree applied on raw formant data (III) set this threshold at 442 Hz, and the percentage of all successfully classified items in this case was 79.81%. Thus, III) is slightly more successful than II), and both II) and III) are much more successful than I). Therefore, our hypothesis that the knowledge of Sg1 would help the classification cannot be supported.

b) Back–front classification

The results of the back–front classification are given in Table 6. Here, methods I) and II) give almost the same overall success ratio, as the optimal threshold of the decision tree is almost 1.0. This means that the optimal separation is reached when the threshold between back and front vowels is the Sg2. As the raw formant-based decision tree (III) gave less accurate results, we can conclude that the frequency normalization to Sg2 improved the classification.

Table 6. Results of the back – front classification

	I (SGR)		II (SGR + decision tree)		III (decision tree)	
	correctly classified	mis-classified	correctly classified	mis-classified	correctly classified	mis-classified
front	91.79%	8.21%	91.53%	8.47%	93.06%	6.94%
back	74.97%	25.03%	74.80%	25.20%	71.45%	28.55%
total	84.95%	15.05%	84.73%	15.27%	84.28%	15.72%
threshold	1.0		0.9999		1344 Hz	

c) Front unrounded non-low – other vowels classification

Table 7. Results of the front unrounded non-low (denoted as "front*") – other vowels classification

	I (SGR)		II (SGR + decision tree)		III (decision tree)	
	correctly classified	mis-classified	correctly classified	mis-classified	correctly classified	mis-classified
front*	20.61%	79.39%	55.19%	44.81%	48.16%	51.84%
other	99.85%	0.15%	97.02%	2.98%	97.30%	2.70%
total	83.17%	16.83%	88.21%	11.79%	86.95%	13.05%
threshold	1.0		0.8778		2018 Hz	

The results of front unrounded non-low – other vowels classification are shown in Table 7. All three methods perform poorly on the separation of front unrounded non-low vowels from other vowels. The weighted averages (in the row labeled "total") show that the best accuracy was reached with method II) (decision tree with F2 normalization to Sg3). There is therefore a small benefit of knowing Sg3, although the optimal threshold after normalization is a somewhat lower frequency than Sg3, as might be expected from the analyses of the vowel plots and formant distributions.

Since there is only one female speaker, the results of method III) can be improved by ignoring this subject. If the decision tree is applied only on the raw formant data of the male speakers, the accuracy is improved by 2.05%, 3.11% and 2.89% (for Sg1, Sg2, and Sg3 separation respectively). Using the data of the male speakers alone, the decision tree-based method (III) yields the highest accuracy in all three cases.

In these classifications, our hypotheses regarding the role of SGRs in dividing the formant space were tested. It was shown that the knowledge of Sg2 and Sg3 may improve the accuracy of automatic vowel classification if using male and female data together. Sg1 was not helpful in automatic classification of spontaneous speech. This is likely due to the fact that the vowels [ɔ] and [ɛ] were considered to be low

vowels. However, in three of the six speakers, they appear to be produced like non-low vowels, with F1 lower than Sg1.

3 Perception test

In this section the results of a perception experiment are described, with the primary aim of testing the hypothesis that the effect of SGRs is observable in speech perception, as well as in speech production. After a short introduction, the methods and goals of the experiment are presented.

3.1. The role of perception in phonology

In the past decades, ample evidence has been presented that speech perception and phonological systems interact (e.g. Hume & Johnson 2001). Phonological contrasts are generally thought to require salient perceptual cues to be present in the acoustic signal. Needless to say, salience is a relative property since the quality and quantity of the cues depends on the context in which the contrast is found. For instance, the place contrast of stops depends largely on the vowel transitions to the flanking vowels, whereas fricative contrasts are typically cued internally. Accordingly, we investigated whether and how the proposed role of SGRs in vowel categorization is utilized in human speech perception.

3.2. Methods and goals

A pilot perception test was designed to investigate the role of the SGRs in human speech perception, based on the measured formant and SGR frequencies as described in Section 2.2. Lulich et al. (2007) investigated a similar question and found that the frequency of Sg2 affects vowel perception, but they used partially synthetic speech as their stimuli. In the present experiment, the feature [+/– back] of two speakers' [ɔ] realizations was chosen as the target for this study. One of them was the female speaker (F1) whose vowel realizations were fairly well separated by her SGRs in accordance with our hypothesis; while the other speaker was a male speaker (M4), whose vowel realizations showed a small vowel space and were therefore not clearly separable on the basis of his SGRs.

Twenty-four [ɔ] vowels were selected from each of the two speakers' spontaneous speech recordings. The F2 values of these samples were distributed in the range 800–1800 Hz (speaker M4) and 1100–1800 Hz (speaker F1). Twenty-nine of them showed a lower F2 than the speaker's Sg2, and the remaining nineteen showed a higher F2. Four or five realizations of several other vowels ([aː, ɛ, o, ø, u]) were also selected from each speaker as distractors. The vowels were isolated with the preceding and following consonant, so that the stimuli were nonsense CVC syllables. Altogether 96 samples were included in the test.

The listeners were 21 native Hungarian females, aged between 20 and 35 years. One of them was excluded from the evaluation because her responses were at chance levels. The task of the listeners was to listen to the samples and decide which vowel they heard. They were told that the stimuli were sound-sequences extracted from spontaneous speech and consisted of CVC-syllables, but not whole words. The stimuli were presented with the Praat program (Experiment MFC function) in

randomized order. Six possible answers (letter symbols of the six vowel phonemes used in the experiment: *a, á, e, o, ö, u*) were displayed on the monitor each time. A response could be given by clicking on one of the symbols. When needed, the sound-sequences could be replayed once. The subjects could take a break after every 24 stimuli. The test was carried out individually using stereo headphones in quiet.

3.3. Results and discussion

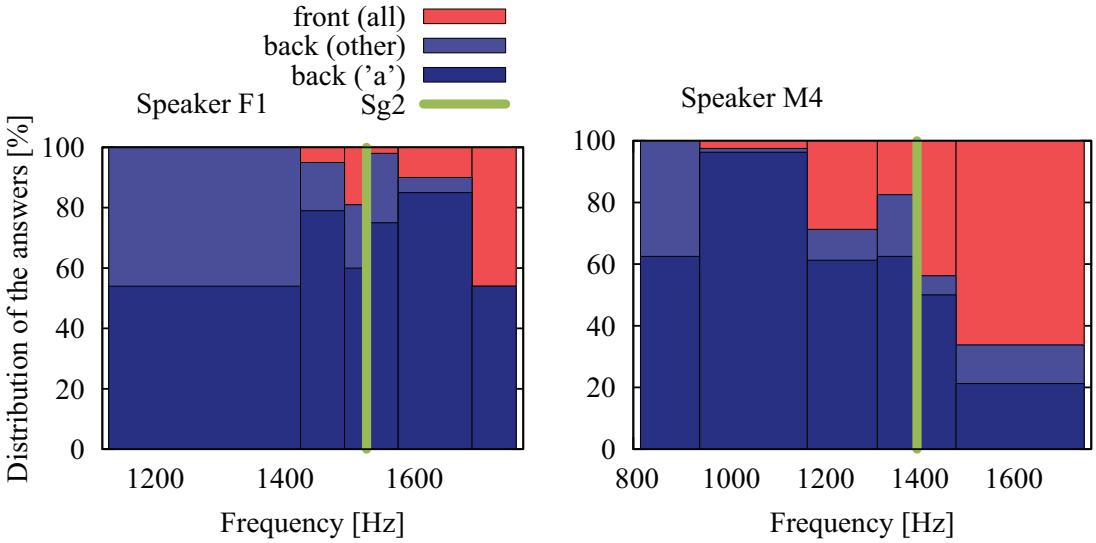


Figure 5. Distribution of perceived vowel categories. Stimuli are represented in groups of four (e.g. the 1100–1410 Hz bar in the left subfigure contains the distribution of answers for four stimuli of Speaker F1 pooled together)

The responses of the subjects in the perception test were sorted into three groups: i) vowels correctly identified as [ɔ], ii) as any other back vowel, or iii) as any front vowel. Figure 5 shows the results of the test. In this figure, each bar represents the responses of the 21 listeners to each of 4 stimuli with F2 frequencies in the range indicated by the width of the bar. The F2 frequencies were measured at the midpoint of the vowel. Back percepts are indicated with dark blue (only vowel [ɔ]) and light blue (back vowels other than [ɔ]), while red color shows the front percepts. The green vertical lines represent the Sg2 of the speakers.

The total back and [ɔ] responses showed a significant negative correlation with the increase of the F2 of the stimuli (female speaker: $r = -0.511$, $p = 0.011$; male speaker: $r = -0.598$, $p = 0.002$), as would be expected. For the female and male speakers (F1 and M4, respectively), 16 and 14 out of the 24 stimuli were recognized correctly by at least 75% of the subjects. The number of [ɔ] and other back vowel responses decreased for F2 higher than Sg2. In the case of the female speaker (F1), only two stimuli were identified as front vowels by most (more than 75%) of the listeners. Both of these stimuli had higher F2 than the speaker's Sg2. Four stimuli of the male speaker M4 were identified by at least 75% of the listeners as front vowels: one had lower F2 than the speaker's Sg2; three had higher F2. Thus, for both

speakers, when at least 75% of listeners agreed that a particular realization of the vowel [ɔ] was a front vowel, that realization did indeed have an F2 higher than Sg2 in 80% of cases.

For speaker M4, an abrupt increase in front percepts is observed when F2 is higher than Sg2. This phenomenon does not clearly occur for speaker F1. A possible explanation for the abrupt jump may be a stronger effect of Sg2 in speaker M4, perhaps because the vowel space has more overlap between vowels than for speaker F1. Another possible alternative is that the F2 happens to be high enough to be perceived as [–back] at about the same frequency as Sg2. Although speaker F1's vowels were clearly separated by the SGRs in nonsense words, her back vowels tended to have F2 frequencies higher than Sg2 in spontaneous speech. It may be that SGRs become more important in perception when other acoustic cues (such as widely distinct formant frequencies, as evidence in the nonsense words of speaker F1) are lacking.

According to these results, the hypothesis that Sg2 affects the perception of the vowel [ɔ] excised from spontaneous speech could be partially but inconclusively supported. For one speaker, perception changed abruptly in the region of Sg2, while for the other speaker this was not observed.

4 General discussion and conclusions

Earlier studies have reported speaker- and phonetic context-dependent effects for the dividing function of Sg2 (Madsack et al. 2008; Jung, 2009b). The present paper showed considerable interspeaker variability for all three of the SGRs: Sg1, Sg2 and Sg3.

We also presented evidence that a style-dependent variability is observable for the dividing function of the SGRs. Furthermore, SGRs may have some effect not only on the production, but also on the perception of Hungarian vowels.

Our results, in accordance with other studies (e.g. Lulich, 2010; Csapó et al. 2009; Jung 2009a; Madsack et al. 2008) confirm that Sg2 is a boundary between front and back vowels for Hungarian, as it appears to be in English, German, and Korean. A difference was found between laboratory speech (isolated nonsense words) and spontaneous speech: the formant spaces and their division by SGRs are generally clearer in nonsense words than in spontaneous speech, as expected. This difference may be due partly to the larger number of vowels analyzed in spontaneous speech than in nonsense words, but it is likely also affected by the difference of the laboratory speech from spontaneous speech per se. As the difference of planning processes for the two speaking styles may lead to a scale-like realization (e.g. for coarticulatory and reduction effects), we considered nonsense words to exemplify hyperspeech, and spontaneous speech to exemplify hypospeech (Lindblom 1990).

In German (Madsack et al. 2008), the low vowel [a] was found to have F1 > Sg1 in most cases, while non-low vowels occurred mostly with F1 < Sg1. Jung (2009a), in a meta-study of formant frequencies reported in the literature, tested the same relation in male and female speakers of ten languages, obtaining similar results: Sg1

plays a role in defining the vowel feature contrast of [low]. In our data, the dividing line between low and non-low vowels is less clear. From the three low Hungarian vowels, only [aː] has its first formant consistently above the first subglottal resonance. The relative place of the vowels [ɔ] and [ɛ] in the formant space depends on the speaker. It must be remembered, however, that Sg1 is more difficult to measure than Sg2 or (given a low enough noise floor) Sg3 in the accelerometer signal. It is also possible that this contributed to our inconclusive results regarding Sg1, though it was not likely a significant source of error.

The effect of Sg3 was found to be relevant in differentiating front unrounded non-low vowels from other front vowels in the study of Csapó et al. (2009) in Hungarian. In the experiments presented here, however, this was only true for the nonsense words of two speakers (M5 and F1), and for only the vowel [iː] in the spontaneous speech of the same speakers plus one additional speaker (M2). The data from other speakers did not support this role for Sg3. It is possible that Sg3 does not have as strong an effect as previously hypothesized. This is not necessarily surprising, since the acoustic coupling between the vocal tract and subglottal airways is expected to decrease at higher frequencies.

The perception test by Lulich et al. (2007) showed that a formant discontinuity near Sg2 influences the perception of backness in the vowel. The stimuli were synthesized utterances with F2 discontinuities at different locations. In this paper, a similar listening test was performed on real speech. C[ɔ]C transitions with different F2 frequencies at the vowel midpoint were extracted from two speakers' spontaneous recordings. The results showed that for one of the tested speakers, an abrupt increase in perceived frontness of the vowel occurred when F2 was higher than Sg2. For the other speaker a similar abrupt increase was not observed. The difference between the two speakers may be related to the size of the vowel space and the overlap between vowel distributions. Another possibility is that vowels with a discontinuity in the formant trajectories due to Sg2 will be more prone to Sg2 effects in perception, but this was not explored in the present study.

Summary

The results of this paper are based on six speakers' speech. In order to further investigate the role of SGRs, more Hungarian speech and accelerometer data need to be collected. This paper presented a preliminary perception test examining the relation of spontaneous speech and SGRs. Our future plans include performing a more fully developed perceptual evaluation.

The results have implications for understanding phonological distinctive features, and applications in automatic speech technologies. The former involves a link between phonological and phonetic interpretation, in the form of supporting some of the claims of Quantal Theory. The latter includes speaker normalization (e.g. Wang et al. 2009) and other related problems in automatic speech recognition. The fact that SGRs are roughly constant for a given speaker may be useful in speaker recognition as well.

5 Acknowledgements

We would like to thank our subjects participating in these experiments and the reviewer for the valuable comments. The research was partially supported by grants from the Hungarian National Scientific Research Foundation (OTKA), No. 78315, from the American National Science Foundation, No. 0905250, by the ETOCOM project (TÁMOP-4.2.2-08/1/KMR-2008-0007) through the Hungarian National Development Agency and the BelAMI project (OMFB-00736/2005) through the Hungarian National Office for Research and Technology, supported by the EU and co-financed by the European Social Fund.

References

Boersma, P. and Weenink, D. 2008. Praat [Computer program] (Version 5.0.20). http://www.praat.org, accessed on 11 Dec 2008.

Beke, A. and Gráczi, T. E. 2010. A magánhangzók semlegesedése a spontán beszédben, [Vowel neutralization in spontaneous speech]. In Navracsics, J. (ed.): *Nyelv, beszéd, írás. Pszicholingvisztikai tanulmányok 1.* [Language, speech, writing. Psycholinguistic studies 1]. Budapest: Tinta Könyvkiadó. 57–65.

Cardillo, G. 2008. ROC curve: compute a Receiver Operating Characteristics curve. http://www.mathworks.com/matlabcentral/fileexchange/19950, accessed on 10 May 2010.

Chi, X. and Sonderegger, M. 2007. Subglottal coupling and its influence on vowel formants. *Journal of the Acoustical Society of America* 122, 1735–1745.

Chomsky, N. and Halle, M. 1968. *The Sound Pattern of English.* Cambridge, MA: MIT Press.

Cranen, B. and Boves, L. 1987. On subglottal formant analysis. *Journal of the Acoustical Society of America* 81, 734–746.

Crosswhite, K. 2004. Vowel Reduction. In Hayes, B., Kirchner, R. and Steriade, D. (eds.): *Phonetically-Based Phonology.* Cambridge University Press. 191–231.

Csapó, T. G., Bárkányi, Zs., Gráczi, T. E., Bőhm, T., and Lulich, S. M. 2009. Relation of formants and subglottal resonances in Hungarian vowels. In: *Proceedings of Interspeech.* 484–487.

Gósy, M. 2008. Magyar spontánbeszéd-adatbázis – BEA [Hungarian spontaneous speech database – BEA]. *Beszédkutatás 2008,* 194–208.

Hayes, B., Kirchner, R. and Steriade, D. 2004. Phonetically-Based Phonology. Cambridge: Cambridge University Press.

Hume, E. and Johnson, K. (eds.) 2001. *The Role of Speech Perception in Phonology.* San Diego: Academic Press.

Ishizaka, K., Matsudaira, M. and Kaneko, T. 1976. Input acoustic impedance measurement of the subglottal system. *Journal of the Acoustical Society of America* 60, 190–197.

Jung, Y. 2009a. *Acoustic articulatory evidence for quantal vowel categories: The features [low] and [back].* Ph.D. thesis. MIT.

Jung, Y. 2009b. Subglottal effects on the vowels across language: Preliminary study on Korean. *Journal of the Acoustical Society of America* 125, 2638.

Kovács, M. 2002. *Tendenciák és szabályszerűségek a magánhangzóidőtartamok produkciójában és percepciójában* [Trends and regularities in the production and perception of vowel length]. Ph.D. thesis. University of Debrecen.

Kovács, M. 2004. Az artikulációs konfiguráció módosulásának kérdése [On the variation of the articulatory configuration]. Presentation on the *VII. Nemzetközi Magyar Nyelvtudományi Kongresszus* [7th International Congress for Linguistics on Hungarian]. Budapest.

Liljencrants, J. and Lindblom, B. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48, 839–862.

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. and Marchal, A. (eds.): *Speech production and speech modelling.* Dordrecht: Kluwer. 403–439.
Lulich, S. M. 2010. Subglottal resonances and distinctive features. *Journal of Phonetics* 38, 20–32.
Lulich, S. M., Bachrach, A. and Malyska, N. 2007. A role for the second subglottal resonance in lexical access. *Journal of the Acoustical Society of America* 122, 2320–2327.
Madsack, A., Lulich, S. M., Wokurek, W. and Dogil, G. 2008. Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs. In: *Proceedings of LabPhon* 11, 91–92.
Mády, K. and Reichel, U. D. 2007. Quantity distinction in the Hungarian vowel systems – Just theory or also reality? In: *Proceedings of ICPhS.* Saarbrücken. 1053–1056.
Mihajlik, P., Révész, T. and Tatai, P. 2002. Phonetic transcription in automatic speech recognition. *Acta Linguistica Hungarica* 49(3–4), 407–425.
Nádasdy, Á. and Siptár, P. 1994. A magánhangzók, [Vowels]. In Kiefer, F. (ed.): *Strukturális magyar nyelvtan 2. Fonológia* [A Structural Grammar of Hungaran 2. Phonology]. Budapest: Akadémiai Kiadó. 42–181.
Ohala, J. J. 1983. The origin of sound patterns in vocal tract constraints. In MacNeilage, P. F. (ed.): *The Production of Speech.* New York: Springer-Verlag. 189–216.
Sjölander, K. and Beskow, J. 2009. *Wavesurfer* [Computer program] (Version 1.8.5). http://www.speech.kth.se/wavesurfer/, accessed on 3 Apr 2009.
Stevens, K. N. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In Denes, P. B. and David, E. E. Jr. (eds.): *Human Communication, A Unified View.* New York: McGraw-Hill. 54–66.
Stevens, K. N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17, 3–45.
Stevens, K. N. 1998. *Acoustic Phonetics.* Cambridge, MA: MIT Press.
Stevens, K. N. and Keyser, S. J. 2010. Quantal theory, enhancement and overlap. *Journal of Phonetics* 38, 10–19.
Wang, S., Alwan, A. and Lulich, S. M. 2008. Speaker normalization based on subglottal resonances. In: *Proceedings of ICASSP.* 4277–4280.
Wang, S., Lulich, S. M. and Alwan, A. 2009. Automatic detection of the second subglottal resonance and its application to speaker normalization. *Journal of the Acoustical Society of America* 126, 3268–3277.
Witten, I. H. and Frank, E. 2005. Using the J4.8 Decision Tree. *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann.