Does vowel recognition relate to pitch?

Dieter Maurer, Christian d'Heureuse, Heidy Leemann
Institute for the Performing Arts and Film IPF, Zurich University of the Arts ZHdK, Switzerland
Inventec Informatik AG, Winterthur, Switzerland

Acoustics Virtually Everywhere
The 179th Meeting of the Acoustical Society of America
7-11 December 2020

Abstract in JASA
Study (PDF online)
Research homepage
Contact: dieter.maurer@zhdk.ch
Version 2021-02-07

Abstract

It was shown that, for vowel sounds, the spectrum relates to fundamental frequency (f_o) and the spectral envelope is ambiguous, often representing two or three different vowel qualities if f_o is varied substantially. Thus, from a speech perception perspective, vowel quality is indicated to relate to the pitch of a vowel sound. In this contribution, a concept is outlined addressing the experimental question of the relationship of vowel recognition and pitch. According to this concept, two or three harmonics are extracted near the spectral peaks of a natural vowel sound, representing assumed *F*1–*F*2 or *F*1–*F*2–*F*3, and the lowest harmonic(s) below the first spectral peak are added. Based on a single extracted harmonics pattern, a series of sounds are synthesised stepwise attenuating the levels of the lowest harmonics with the aim to effect a low-to-high transition of the highest common harmonics factor and pitch. Pitch level and vowel recognition of the sounds is then investigated by means of a listening test. – Results of a first pilot test based on sounds of mid-closed vowels /e, ø, o/ confirmed the vowel–pitch relation hypothesis and also revealed cases of sounds with double-pitch and double-vowel recognition.

BACKGROUND

The vowel spectrum of natural sounds relates to fundamental frequency (f_o).

Examples

References

The spectral peak patterns, the estimated formant patterns and the spectral envelopes of vowel sounds are ambiguous, often representing two or three different vowel qualities if f_o is varied.

Examples

References

In vowel synthesis based on series of equal amplitude harmonics, the recognised vowel quality can be changed by either altering the frequency of a single lower amplitude harmonic or only changing the frequency spacing of the higher harmonics. In both cases, the change in vowel quality is indicated to relate to the highest common factor of the harmonics and to pitch.

Examples

References

The recognition of sinewave vowel sounds is also indicated to relate to pitch.

Examples

References

The relation between recognised vowel quality and the vowel spectrum is nonuniform.

Examples

References

THESIS

Therefore, from a speech perception perspective, two indications exist:

> **The recognised vowel quality relates to the pitch of a vowel sound.**
>
> **This relation is nonuniform.**

Here, the first aspect of the pitch-relation is addressed.

GENERAL APPROACH AND PROCEDURE

Creation of a model synthesis experiment

The following model synthesis experiment was designed in order to address the vowel–pitch relation hypothesis.

Firstly, natural vowel sounds are to be selected for which

- the vowel-related spectral peaks are represented by dominant or prominent harmonics, *D*1–*D*2 (back vowels and /a/) or *D*1–*D*2–*D*3 (front vowels);
- these harmonics relate to formant patterns *F*1–*F*2 or *F*1–*F*2–*F*3 as commonly estimated for the sounds in question;
- the first dominant or prominent harmonic *D*1 is above the fundamental *H*1 (*D*1 > *H*1); the higher dominant or prominent harmonic(s) are integer multiples of the first dominant or prominent harmonic.

Secondly, the dominant or prominent harmonics and also the low harmonic(s) below *D*1 are to be extracted.

Thirdly, harmonic synthesis is to be applied with

- the patterns of the dominant or prominent harmonics keept unchanged;
- the low harmonic(s) below *D*1 stepwise attenuating until this harmonic is or these harmonics are deleted.

Fourthly, vowel and pitch recognition is to be tested.

For illustration, see below, method section, Figures 1 to 3.

Special character of the synthesised sounds

Sounds that are synthesised on the basis of these patterns of dominant or prominent harmonics (hereafter *D*-patterns) have special characteristics:

- The highest common factor HCF of the pattern including the harmonic(s) below *D*1 is lower than HCF of the pattern excluding the harmonic(s) below *D*1;
- therefore, the periodicity of the two types of sounds differ, affecting pitch perception and recognition;
- however, the spectral energy maxima are kept unchanged for both types of sounds, and if formant patterns are estimated, they also will prove to be unchanged;
- there is no existing concept of spectral envelope estimation accounting for sounds of this kind and their spectral representation of vowel quality.

Dissociation of fundamental frequency and pitch

In series of synthesised sounds of this type related to a single natural reference sound, above all in the course of the attenuation of the low harmonic(s) below D1, two concurrent HCF and, accordingly, **two concurrent sound periodicities occur to which perception and recognition can refer to:**

- HCF including the lower harmonic(s)
- HCF of the dominant harmonics only

However, only one fundamental frequency is usually attributed to a single sound.

Expectation

Cases of synthesised sounds are expected to occur for which

- the pitch of the sounds produced with both the dominant or prominent harmonics and the low harmonic(s) below *D*1 is recognised as lower than the pitch of the sounds produced with the dominant or prominent harmonics only;
- in parallel, the recognised vowel quality of the first sounds mentioned differs from the quality of the second sounds;
- above all in the course of attenuating the lower harmonic(s) below *D*1, two pitches and/or two vowels may be recognised.

General Procedure

In general, the following procedure is proposed.

Create a large sample of vowel sounds: To find sounds fulfilling the above conditions of dominant or prominent harmonics being associated with spectral peaks, the creation of a large sample of vowel sounds is needed. (For our basis of the Zurich Corpus, see next text box.)

Select natural vowel sounds with a specific spectral peak pattern: In the sample created, search for natural sounds for which the vowel-related spectral peaks are associated with dominant or prominent harmonics *D*1–*D*2 (sounds of back vowels and of /a/) or *D*1–*D*2–*D*3 (sounds of front vowels), *D*1 being above *H*1, and *D*2 or D2 and *D*3 being integer multiples of *D*1.

Extract the *D*-pattern of a sound spectrum and add lower harmonics < *D*1: Extract the dominant or prominent harmonics and also the low harmonic(s) below *D*1 from a selected natural sound.

Perform harmonic synthesis with stepwise attenuating the low harmonic < *D*1: Based on the extracted *D*-pattern and the low harmonic(s), perform harmonic synthesis and attenuate the level(s) of the low harmonic(s) step by step.

Perform vowel and pitch recognition tests: Test vowel and pitch recognition for the synthesised sounds, either separately or jointly.

Note

The spectral representation of vowel quality is nonuniform (see Background). Therefore, results may be dependent on vowel qualities and on f_o levels of the natural reference sounds investigated as well as on respective extracted *D*-patterns (see also the Discussion).

PILOT STUDY – METHOD

On the basis of the general experimental design, a pilot study was conducted applying the following method.

Sounds investigated

On the basis of the Zurich Corpus, sounds of the three long Standard German vowels /e, ø, o/ produced by female speakers at f_o of c. 220–250 Hz were investigated.

Visit the Zurich Corpus

Sound selection

Examples of natural vowel sounds produced as monophthongs at f_o of c. 220–250 Hz were selected which fulfilled three conditions:

- The vowel-related spectral peaks were associated with dominant harmonics; however, for sounds of /o/, a second spectral peak indicated by only a prominent harmonic was also accepted
- The first dominant harmonic is not the fundamental *H*1
- The higher dominant (or prominent) harmonic(s) are integer multiples of the low dominant harmonic

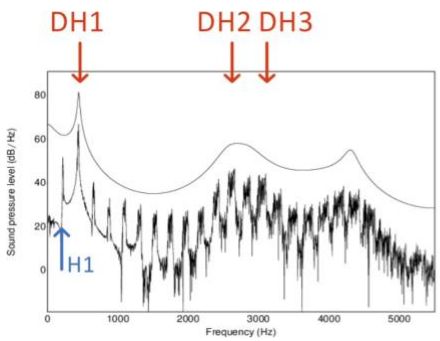

Figure 1: Vowel spectrum of a natural sound of /e/ produced in V context at *f*o of c. 200 Hz.

The first three spectral peaks are associated with dominant harmonics D1–D2–D3 = c. 440–2640–3080 Hz.

Play black this sound

For each of the three vowels /e, ø, o/, two natural sounds were selected fulfilling these conditions. Two peaks for /o/ and three peaks for /e, ø/ were considered vowel related.

Extraction of harmonic patterns

Subsequently, dominant (or prominent) harmonics *D*1–*D*2 or *D*1–*D*2–*D*3 were extracted (including their actual levels and dynamic contours), and the first harmonic *H*1 of the natural sound below *D*1 was added.

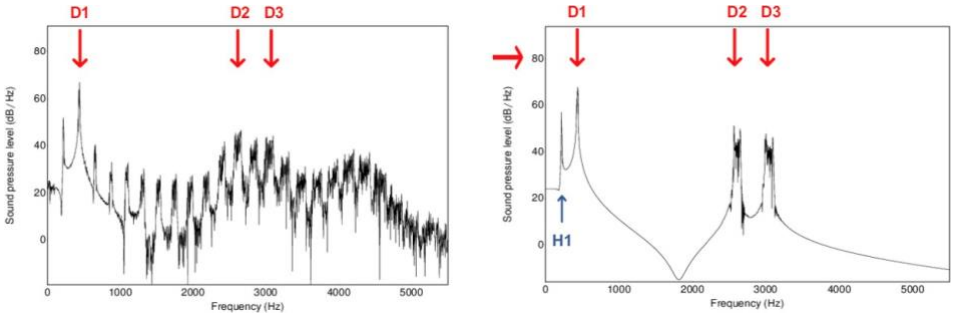

Figure 2: Natural sound of /e/ produced at f_o of c. 220 Hz, with dominant harmonics *D*1–*D*2–*D*3 = c. 440–2640–3080 Hz, and extracted pattern *D*1–*D*2–*D*3 with added *H*1 = c. 220 Hz.

- Play black this sound pair
- See all six pairs of natural vowel sounds and extracted harmonics
- Option: See also formant pattern ambiguity for the six natural reference sounds

Harmonic synthesis

For all six configuration of extracted harmonics, vowel sounds were synthesised with step by step attenuation of *H*1.

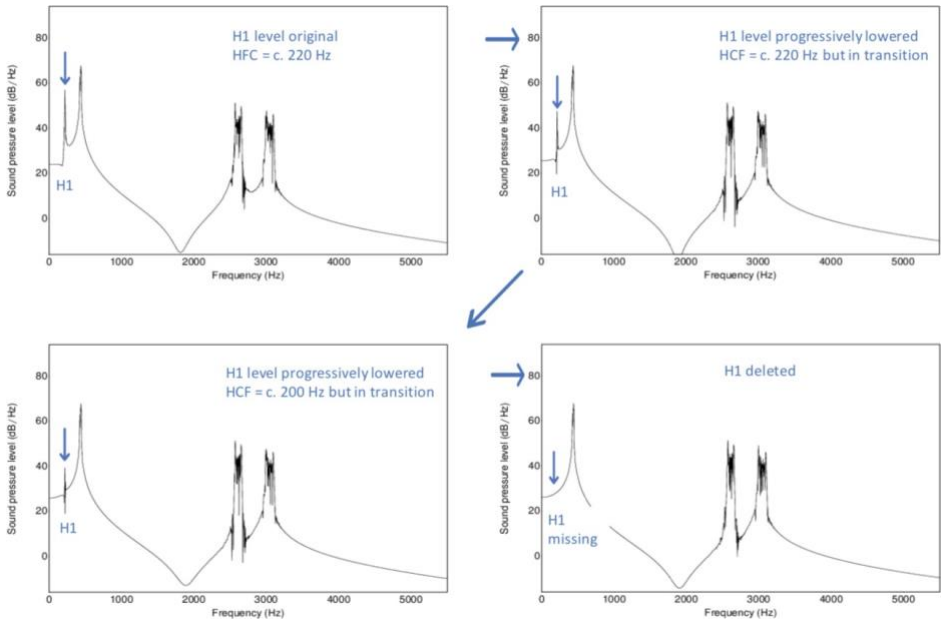

Figure 3: Illustration of *H*1 attenuation, with the *H*1 levels shown = 0–10–20–100 dB.

Eight attenuation levels of *H*1 were investigated: 0/-5/-10/-15/-20/-30/-50/-100 dB

Play black these four sounds

Tool

Extraction of the harmonics and resynthesis was conducted using the HarmSyn harmonic synthesiser.

Note that, using this tool, the dynamic course of the harmonics can be analysed and resynthesised including frequency and level variation of the harmonics over time and including harmonic level attenuation.

Reference

Sound sample created

For each single harmonic configuration, eight synthesised sounds were produced related to the eight attenuation levels of *H*1 mentioned of: 0/-5/-10/-15/-20/-30/-50/-100 dB

A sample of 48 single sounds in total was created for further investigation.
See all sound series, and listen to the sounds

Vowel and pitch recognition tests

Five experienced listeners (professionally trained singers) participated in the below listening tests. (Note that these listeners were also involved in a large number of listening tests for the Zurich Corpus; see above for more information on the Corpus.)

Testing vowel and pitch recognition of opposing sounds (without and with full attenuation of *H*1)

Vowel and pitch recognition of sounds with unchanged *H*1 and with fully deleted *H*1 were tested in three subtests.

In these three subsets, for each test item, the listeners were asked to assign only one vowel (prominent or dominant vowel) or only one pitch (prominent or dominant pitch).

In the first subtest, each test item consisted of a single sound and the listeners were asked to assign a vowel quality (forced choice, all long Standard German vowels and schwa).

In the second and third subtest, each test item consisted of two synthesised sounds (with a 1 sec pause in between): a sound with an unchanged *H1* configuration versus a second sound with fully deleted *H1*, and vice versa (AB and BA order). – In the second subtest, the listeners were asked to assign a vowel quality to the second sound presented (forced choice, see above). – In the third subtest, the listeners were asked to label the pitch difference between the two sounds as either "falling", "rising" or "flat" ("flat"=no pronounced pitch difference).

Testing vowel and pitch recognition of opposing and transitional sounds

Vowel and pitch recognition of all sounds was tested in three subtests.

In these three subsets, for each test item, the listeners were again asked to assign only one vowel (prominent or dominant vowel) or only one pitch (prominent or dominant pitch).

In the first subtest, each test item consisted of a single sound with an attenuated *H1* configuration and the listeners were asked to assign a single vowel quality (forced choice; see above).

In the second and third subtest, each test item consisted of two synthesised sounds (with a 1 sec pause in between): a sound with an unchanged *H1* configuration versus a second sound with attenuated or deleted *H1* configuration or, inversely, a sound with deleted *H1* configuration versus a second sound with unchanged or attenuated configuration (AB and BA order). – In the second subtest, the listeners were asked to assign a vowel quality to the second sound presented (forced choice, see above). – In the third subtest, the listeners were asked to label the pitch difference between the two sounds as either "falling", "rising" or "flat" ("flat"=no pronounced pitch difference).

Testing double-vowel and double-pitch recognition of opposing and transitional sounds

In the fourth subtest, each test item consisted of a single sound and the listeners were asked whether they recognise one or two vowel qualities.

Likewise, in the fifth subtest, each test item consisted of a single sound and the listeners were asked whether they recognise one or two pitches.

PILOT STUDY – RESULTS

Vowel and pitch recognition of opposing sounds (without and with full attenuation of *H*1)

Table 1. Recognition results of pitch and vowel quality tested for opposing sounds of a D-pattern without and with full attenuation of H1 (vowel recognition = 2 subtests with 10 identifications, pitch recognition = 1 test with 5 identifications; see Method section). Columns 1–3 = sound production (S=series number, fo=fundamental frequency in Hz, V=vowel quality intended). Columns 4–8 = harmonic configuration of synthesis (H1=fundamental, AH1=attenuation of H1, D(i)=dominant harmonics, approximate values in Hz). Columns 9–14 = recognised vowel qualities (total of both tests). Columns 15–16 = recognised pitch level for the comparison of sounds related to a D-pattern without and with full attenuation of H1. – Recognition rates ≥ 80% of mid-closed vowels and/or of low pitch are highlighted in blue, and recognition rates ≥ 80% of closed vowels and/or of high pitch are highlighted in red.

Natural sound			Harmonic synthesis					Vowel recognition						Pitch recognition	
S	f_o	V	H1	AH1	D1	D2	D3	o	ø	e	u	y	i	low	high
1	220	e	220	0	440	2640	3080			10				5	
			–	-100									10		5
2	250	e	250	0	500	2500	3000		1	9				5	
			–	-100						1			9		5
3	220	ø	220	0	440	1760	2640		10					5	
			–	-100								10			5
4	220	ø	220	0	440	1760	2640		10					5	
			–	-100					1			9			5
5	220	o	220	0	440	880	–	10						5	
			–	-100							10				5
6	220	o	220	0	440	880	–	8			2			5	
			–	-100							10				5

The results of the vowel and pitch recognition tests for opposing sounds of a *D*-pattern without and with full attenuation of *H*1 (test details, see Method section) showed that

- no attenuation of *H*1 was associated with a lower pitch level and a mid-closed vowel quality;
- full attenuation of *H*1 (*H*1 deleted) was associated with a higher pitch level and a closed vowel quality.

See this table in PDF format

See all opposing sounds, and listen to the sounds

Vowel and pitch recognition of opposing and transitional sounds

Table 2. Recognition results of pitch and vowel quality tested for all sounds of a D-pattern with stepwise attenuation of H1 from 0 db to -100db (vowel recognition = 2 tests with 15 identifications, pitch recognition = 1 test with 10 identifications; double-vowel recognition = 1 test with 5 identifications; double-pitch recognition = 1 test with 5 identifications; values marked with "*" are taken from Table 2; details see Method). Columns 1–3 = sound production (S=series number, fo=fundamental frequency in Hz, V=vowel quality intended). Columns 4–8 = harmonic configuration of synthesis (H1=fundamental, AH1=attenuation of H1, D(i)=dominant harmonics, approximate values in Hz). Columns 9–14 = recognised vowel qualities (total of both tests). Columns 15–16 = recognised pitch level for the comparison of sounds related to a D-pattern with and without full attenuation of H1. – Recognition rates ≥ 80% of mid-closed vowels and/or of low pitch are highlighted in blue, and recognition rates ≥ 80% of closed vowels and/or of high pitch are highlighted in red.

Natural sound			Harmonic synthesis					Vowel recognition						Pitch recognition	
S	f_o	V	H1	AH1	D1	D2	D3	o	ø	e	u	y	i	low	high
1	220	e	220	0	440	2640	3080			13			2	10	
			(200)	-5						10			5	9	1
			(200)	-10						11			4	9	1
			(200)	-15						6			9	9	1
			(200)	-20						5			10	7	3
			(200)	-30									15		10
			(200)	-50									15		10
			–	-100									15		10
2	220	e	220	0	440	2640	3080		1	13			1	10	
			(200)	-5					2	13				10	
			(200)	-10					1	13			1	10	
			(200)	-15						13			2	7	3
			(200)	-20						9			6	6	4
			(200)	-30						9		2	4	6	4
			(200)	-50						3			12		10
			–	-100						3			12		10
3	220	ø	220	0	440	2640	3080		15					10	
			(200)	-5					15					10	
			(200)	-10					14			1		10	
			(200)	-15					13			2		10	
			(200)	-20					12			3		8	2
			(200)	-30					8			7		5	5
			(200)	-50								15			10
			–	-100								15			10
4	220	ø	220	0	440	2640	3080		15					10	
			(200)	-5					15					10	
			(200)	-10					14			1		10	
			(200)	-15					11			4		8	2
			(200)	-20					9			6		6	4
			(200)	-30					7			8		5	5
			(200)	-50					1			14			10
			–	-100					1			14			10
5	220	o	220	0	440	2640	3080	15						10	
			(200)	-5				14			1			10	
			(200)	-10				12			3			9	1
			(200)	-15				10			5			7	3
			(200)	-20				7			8			6	4
			(200)	-30				6			9			4	6
			(200)	-50							15				10
			–	-100							15				10
6	220	o	220	0	440	2640	3080	12			3			10	
			(200)	-5				12			3			10	
			(200)	-10				12			3			9	1
			(200)	-15				10			5			9	1
			(200)	-20				8			7			8	2
			(200)	-30				6			9			6	4
			(200)	-50							15				10
			–	-100							15				10

See this table in PDF format

Referring to the columns 9–14 and 15–16 in Table 2, the results of the vowel and pitch recognition tests for all sound series of a *D*-pattern with stepwise attenuation of *H*1 from 0 to -100 dB showed that

- marked between-speaker differences occurred for both vowel and pitch recognitions in the transitions from the oppositions of mid-closed vowels and lower pitch (no attenuation of *H*1) to closed vowels and higher pitch (full attenuation of *H*1).

See all sound series, and listen to the sounds (link is also given in the Method section)

Double vowel and double pitch recognition

Referring to the Columns 17 and 18 in Table 3 (extension of Table 2, see link below), the results of testing double vowel and double pitch recognition showed that

- numerous cases of sounds occurred for which two vowel qualities and/or two pitches were recognised.

See Table 3 in PDF format (extension of Table 2)

Further details of the results

Details of all pitch and vowel recognition results are given in separate documents.

See the details of the recognition results

DISCUSSION

In the present experiment, two dominant or prominent harmonics *D*1–*D*2 (sounds of /o/) or three dominant harmonics *D*1–*D*2–*D*3 (sounds of /e, ø/) of the spectra of natural mid-closed vowel sounds produced at f_o in the range of c. 220–250 Hz were extracted, the dominant or prominent harmonics corresponding to formant frequency patterns *F*1–*F*2 or *F*1–*F*2–*F*3 commonly estimated for these natural sounds. Then, the fundamental *H*1 of the sound spectrum in question (the harmonic below the first spectral peak) was added. Based on this type of harmonic patterns, series of sounds were synthesised stepwise attenuating the levels of *H*1 with the aim to trigger a low-to-high transition of the highest common harmonics factor and of the recognised pitch. Finally, vowel and pitch recognition of the sounds was investigated.

Firstly, **the results support the vowel–pitch relation hypothesis**, here in terms of a pronounced general tendency for a mid- closed to closed vowel quality shift associated with an increase of the recognised pitch.

Secondly, and most importantly, **cases of sounds with double-vowel and/or double-pitch recognition occurred**, a manifestation which strongly underpins the vowel–pitch relation hypothesis.

However, within the transitional phase between the low-pitched sounds of mid-closed vowels and the high-pitched sounds of closed vowels, **the pitch and vowel shifts did not obey a strict parallelism**: Vowel shifts occurred before pitch shifts, and vice versa.

The new indication of vowel quality recognition as related to pitch offers a coherent explanation of earlier findings that the vowel spectrum of natural sounds is related to f_o and that formant patterns and spectral shapes are ambiguous representations of vowel quality as well as of the results of the vowel synthesis experiments referred to in the Background section.

All of the above indications that vowel quality recognition is related to pitch stand in strong contrast to the claim of the prevailing acoustic theory of the vowel that either the formant pattern or the spectral envelope acoustically represents vowel quality (see Hillenbrand and Houde, 2003; Swanepoel et al., 2012): In the present experiment, estimated formant patterns do not differ for the vowel quality shifts of the sounds observed, and current concepts of spectral envelope estimation do not account for the harmonic configurations of the vowel spectra discussed here.

As one of the consequences, the spectrograms of two sounds recognised as two different vowels do not indicate this difference, as is illustrated here in the comparison of the first sound pair of /e/ and /i/ shown in Table 1.

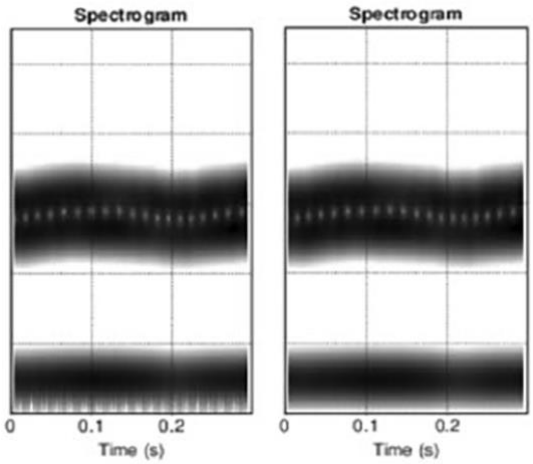

Figure 4: Sound spectrogram of /e/ (left) and /i/ (right) of the sound pair shown in Series of Table 1.

In order to further investigate and clarify the role of pitch for vowel quality recognition and its impact on prevailing acoustic theory, further research on the matter is needed including the development of improved experimental designs.

With respect to such future research, special attention should be given to the fact that, in the experiments referred to, the general indication of a vowel–pitch relation proved to be robust but a strict parallelism vowel and pitch recognition was lacking. Special attention should also be given to the many indications of a nonuniform relation between spectral characteristics and vowel quality (see Background).

Concerning the nonuniform spectral representation of vowel quality, when designing the present experiment, we have experienced that the sound quality of this type of synthesis depends on the vowel quality and on f_o of the natural reference sounds and that this dependence affects the inter-listener accordance or confusion of both vowel and pitch recognition. For similar experimentations, we recommend to first investigate natural reference sounds of mid-closed vowels produced at f_o of 200–250 Hz because their lower vowel spectrum was found to strongly relate to an increase in f_o and, consequently, single formant patterns as well as single spectral envelopes of these sounds are in most cases ambiguous in that they represent mid- closed and closed vowel qualities (see Background). Subsequently, the investigation may be extended to sounds of other vowel qualities produced at lower or higher f_o.

References
Miscellaneous / Corrections

ACKNOWLEDGEMENT

This work was supported by the Swiss National Science Foundation SNSF, Grant 100016_159350.

AUTHOR INFORMATION

Dieter Maurer, Prof. Dr.
Institute for the Performing Arts and Film IPF, Zurich University of the Arts ZHdK, Switzerland; dieter.maurer@zhdk.ch

Christian d'Heureuse
Inventec Informatik AG, Winterthur, Switzerland; chdh@inventec.ch

Heidy Suter
Institute for the Performing Arts and Film IPF, Zurich University of the Arts ZHdK, Switzerland; heidy.suter@zhdk.ch