Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Deviations of perceived vowel quality as a result of F3 manipulations

Thorsten Smit, Friedrich Türckheim, Andreas Jakob and Robert Mores

Department of Technology, University of Applied Sciences, Hamburg, Germany

PACS: 43.71.Bp, 43.71.Gv

ABSTRACT

This paper investigates impacts of $F3$ manipulations within given human voice signals. For this purpose two psychoacoustic experiments were carried out. Following the source filter theory of speech production, two modifications to the formant $F3$ have been investigated, the impact of shifting frequency and of widening bandwidth on perceived vowel quality. Parametric formant manipulations are possible by using the method of linear prediction (LP) analysis, root extraction of LP data and FIR zero-pole design. For the sake of control, test sounds are pitched synthetically. As reference for psychoacoustic tests natural voice signals were used across a wide range of the vowel quality. Subjects had to rate similarity of perceived vowel quality of two manipulated sounds against the original reference sound. A general result from the study is that vowel quality perception is rather tolerant against bandwidth manipulations but quite sensitive to frequency manipulations. Only 60% of the subjects perceived vowel quality dissimilarities even when the bandwidth of $F3$ had been increased by about 600 Hz. On the contrary, $F3$ frequency shifts even as low as 150 Hz already evoked likewise perceptual differences for 80% of the subjects.

INTRODUCTION

Formants are commonly divided into speaker dependent characteristics $F3$, $F4$ and vowel characteristic formants $F1$, $F2$ [1, 2, 3]. Precise vowel classification results can be achieved by combining pitch ($F0$), $F1$ and $F2$ in a closed analytical polynomial, verified against the vowel trapeze of the International Phonetic Association (IPA) [4]. It is commonly believed, that even front vowels like [i] or [e] can be clearly classified without higher formants, as shown in experiments [4]. However, the early work of Stumpf on speech fundamentals would not suggest such simplification [5].

Other studies show that there is a robust speaker dependent link between both, the lower ($F1$, $F2$) and higher ($F3$, $F4$) formants [6, 7]. In that, it seems that the formant structure of human speech follows some functional regularities. It also has been shown that $F3$ is predictable by $F1$ and $F2$. The possibility to predict $F3$ by means of the lower formants can be regarded as specific to the human vocal tract.

The motivation of this study was to investigate perception dependencies by varying the formant structure synthetically. For this purpose, this study aims at verifying tolerable frequency shifts and bandwidth widenings. The findings would possibly allow to model mentioned functional regularities of $F3$ with $F1$ and $F2$, and to further qualify vowel quality extraction.

Psychoacoustic investigations were carried out in order to compare synthesized and naturally spoken vowels. For this purpose, synthetic vowel samples were pitched at $F0$ of the respective natural vowel and subsequently filtered by $F3$ manipulated transfer functions.

This paper is organized as follows: in the section *Psychoacoustical Tests* the test method, the stimuli and subject informations will be discussed. The second section contains the test results. Finally, the last two sections will show some application possibilities of the upcoming results.

PSYCHOACOUSTIC TESTS

In the present study two psychoacoustic listening tests ("paired comparison tests") were carried out in which subjects had to compare naturally spoken reference vowels with synthetically pitched vowels.

In the test, all sounds were accessible by individual play buttons, and the subjects were free to choose any starting point, sequence or number of repetitions for their comparison task.

After each test, the subjects had to answer the question "Which of the synthetic vowels was more similar to the naturally spoken vowel? Please decide on the basis of a change in vowel quality!" . There were three possible answers DEC 1: "clear decision for one of the synthetic vowels", DEC 2: "no decision, both synthetic vowels had the same vowel quality as the natural vowel" and DEC 3: "no decision, both synthetic vowels had an equally large difference to the natural vowel".

Prior to the the test a training was accomplished to ensure that the subjects clearly understood the meaning vowel quality. Further, the test conductor demonstrated potential vowel quality changes and the extent of possible perception drifts.

Stimuli

In this present study five exemplary vowels, taken from *IPA help 2.1, SIL International* which is available at [8], were chosen as naturally spoken vowel samples. These natural vowels were taken to cover a wide range of vowel qualities and tongue positions (for details see Fig. 1).

As noted above, the following signal processing is based on LP analysis. The mathematical background is briefly reviewed to understand the upcoming manipulations of individual formants.

The autoregressive (AR) model of LP analysis is given by

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} = \frac{1}{A(z)} \qquad (1)$$

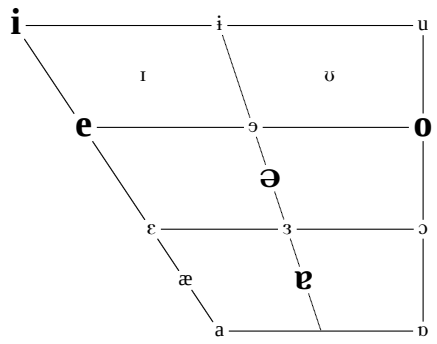

Figure 1: IPA vowel trapeze, reference sounds for the test are marked with bold letters

where $z = r \cdot \exp(-j\omega/F_s)$ is a complex number with magnitude r and angle ω/Fs and p is the LP-order which depends on the sampling frequency F_s. LP analysis computes the coefficients a_k.
With r = 1 we follow the unit circle of the complex z-plane and $H(z)$ can be simplified to $H(\omega)$. Its power spectrum $|H(\omega)|^2$ contains peaks at formant frequencies. The corresponding magnitudes are inversely related to their bandwidths. These resonances occur at radial frequencies where roots of $A(z) = 0$ are crossed while passing through the unit circle (see Fig.). It becomes evident that it is very useful to transform the complex polynomial $A(z)$ into the factorized form. It is given by

$$A(z) = 1 - \sum_{k=1}^{p} a_k \cdot z^{-k} = z^{-p} \cdot \prod_{k=1}^{p} (z - \psi_k). \qquad (2)$$

where each root pair ψ_k can be written as a complex number $\psi_k = r_k \cdot \exp(\pm j\theta_k)$.
Now, starting from this zero-pole representation form, it is easy to evaluate the center frequencies and the 3-dB bandwidths of all formants [9]. These relationships can be expressed by

$$f_k = \frac{F_s}{2\pi}\theta_k \qquad (3)$$

$$B_k = \frac{F_s}{\pi}\ln\left(\frac{1}{r_k}\right) \qquad (4)$$

$$\text{with } \theta_k = \tan^{-1}\left(\frac{\text{Imag}\,\{\psi_k\}}{\text{Real}\,\{\psi_k\}}\right) \text{ and } r_k = |\psi_k|$$

for the k-th formant.
If the computation time does not matter, equation (2) can be solved by standard root solvers.
The separation of *formant roots* and *less significant roots* is commonly carried out by the use of bandwidth criteria which are equivalent to minimum radius criteria. With (4) it can be expressed by

$$r_{k,min} = \exp\left(-\frac{B_{k,max} \cdot \pi}{F_s}\right), \qquad (5)$$

The literature suggests a maximum bandwidth $B_{k,max}$ of 150 Hz, 200 Hz, and 300 Hz for the first three formants, respectively [10].

F3 manipulation by root shifting - F3 manipulations were carried out by shifting the corresponding roots, ψ_{F3}.
Such shifting can be parameterized by displacements of the absolute value $|\psi_{F3}| = r_{F3}$ (bandwidth manipulation) and separately by displacements of the angle $arg\{\psi_{F3}\} = \theta_{F3}$ (center frequency manipulation). This results in manipulated polynomials $A_m(z)$ and in manipulated transfer functions $H_m(z)$.
The set of used *F3* manipulations for each natural vowel can be found in Tab. 1 and it's spectral impacts are shown in Fig. 2.

No.	Frequency shifts in Hz	Bandwidth extensions in Hz
1	-300	300
2	-150	600
3	150	1000
4	300	

Table 1: $F3$ root manipulation values

Permutations cover the entire range of possible pairwise comparisons, six pairs for frequency manipulations and three pairs for bandwidth manipulations for each natural vowel.

Subjects and test environment

A total of 41 german-speaking persons attended the listening tests of this study. None of the subjects suffered from hearing impairments. Furthermore, all subjects claimed not to be experienced in terms of speech intelligibility or psychoacoustic tests. The age of the participants ranged between 17 to 53 years of age.
Subjects were post selected by individual reliability, individual disagreements [11] and by individual rating times.
In summary, only one subject had to be sorted out.
All tests were held in well-defined environments. The mono sounds were presented via headphones from a PC and subject's decisions were recorded directly on the PC.

RESULTS

The analysis of raw data bases on a simple rating scheme. It was organized as follows $DEC\ 1 = 1\ pt$, $DEC\ 2 = 0.5\ pt$ and $DEC\ 3 = 0.25\ pt$.
For clarity the reference vowels in Fig. 3 + 4 were marked.
The key finding of this study is: human perception is far more sensitive to formant frequency shifts than to bandwidth widening. Even a slight frequency shift of $F3$ ($\geq$ 150 Hz) evokes changes of perceived vowel quality, whereas only a relatively large bandwidth widening of $F3$ ($\geq$ 600 Hz) evoke likewise changes of perceived vowel quality.
A more detailed evaluation shows that only 60% of the subjects were able to perceive changes of the vowel quality when bandwidth widened by more than 600 Hz. In contrast, frequency shifts of 150 Hz were perceived by 80% of the subjects.

Frequency shifting results

The distributions in Fig. 3 reveal the following results:

- Perceived vowel quality is dependent on the consistent position of $F3$.
- Larger frequency manipulation of $F3$ evoke larger drifts of perceived vowel quality.
- There is no symmetrical trend, dimension or weighting for perceptual changes.
- $F3$ manipulations have stronger impact on the perception of front vowels than on the perception of middle and back vowels.

Bandwidth extension results

The distributions in Fig. 4 reveal the following results:

- Manipulation of $F3$-bandwidth has stronger impacts on perceived vowel quality of front vowels than on middle and back vowels.
- Only a bandwidth widening beyond 600 Hz cause significant changes of the vowel quality perception. Further increase of the bandwidth has almost no impact on

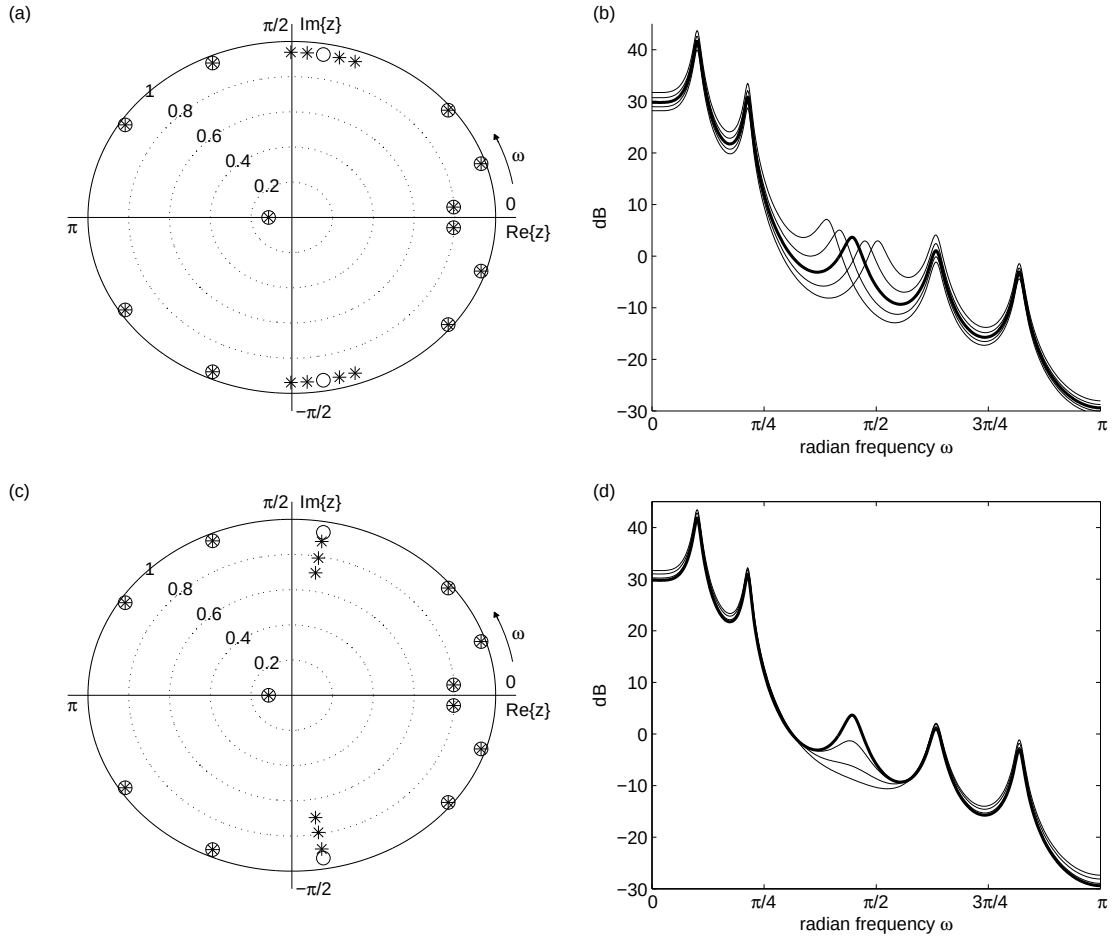

Figure 2: Fig.(a),(c) = poles of $H(z)$ and $H_m(z)$, circles: poles of $H(z)$, stars: poles of $H_m(z)$
Fig.(b),(d) = corresponding spectra $|H(\omega)|_{dB}$ and $|H_m(\omega)|_{dB}$ in dB, bold line = $|H(\omega)|_{dB}$, thin lines = $|H_m(\omega)|_{dB}$
see Fig.(a),(b) for frequency shifts, see Fig.(c),(d) for bandwidth extensions

perception.

- Remarkable is the gradient of the distribution in Fig.4(d) of the vowel [*o*], which stands in great contrast to the other vowels.
- Middle vowels seem to be robust against bandwidth extensions of $F3$.

DISCUSSION

There are two key aspects which should be discussed in this section:

1. How can the results of this present study be utilized?
2. Are there tolerance limits for $F3$ displacements?

1. - As already mentioned in the introduction, vowel classifications are already precise when using $F0$, $F1$ and $F2$ only. Nevertheless, classifications inaccuracy and classification gross errors occur. These are caused by incorrect formant extractions. So far there are no subsequent control mechanisms to avoid such faulty extractions. For this reason, a post-screening of the vowel classification is proposed. Therefore, the functional regularity of the formant constellation, as mentioned above [6, 7], can be used to avoid faulty extractions. $F3$ prediction allows for comparing prediction and determined results of $F3$ for correctness.

2. - Boundaries in which F3 may vary without effecting on the perceived vowel quality are specified separately in Tab. 2.
Note:

Based on the results of this study, the proposed $F3$-prediction-control-mechanism for vowel classification should be most focused on the frequency bands for prediction than on bandwidth criteria. This is due to the fact that $F3$ bandwidth variations seem to be rather negligible.

Frequency boundaries	Maximum 3-dB bandwidth
± 150 Hz	$B_{F2} + 600$ Hz

Table 2: Bandaries of a meaningful $F3$ prediction, B_{F2} = bandwidth of formant $F2$

CONCLUSION

This paper presented drifts of perceived vowel qualities caused by $F3$ manipulations. It was shown that human perception is quite sensitive to frequency manipulations but not very sensitive to bandwidth manipulations of $F3$. This was exposed by the results of psychoacoustic tests in which subjects had to rate vowel quality differences of synthetically manipulated vowels to naturally spoken vowels. Accordingly, only 60% of the subjects were able to notice differences between bandwidth manipulated sounds and reference sounds while 80% of the subjects perceived frequency manipulations.
In agreement with earlier findings [5], it can be confirmed that the perception of front vowels is dependent on higher formants,

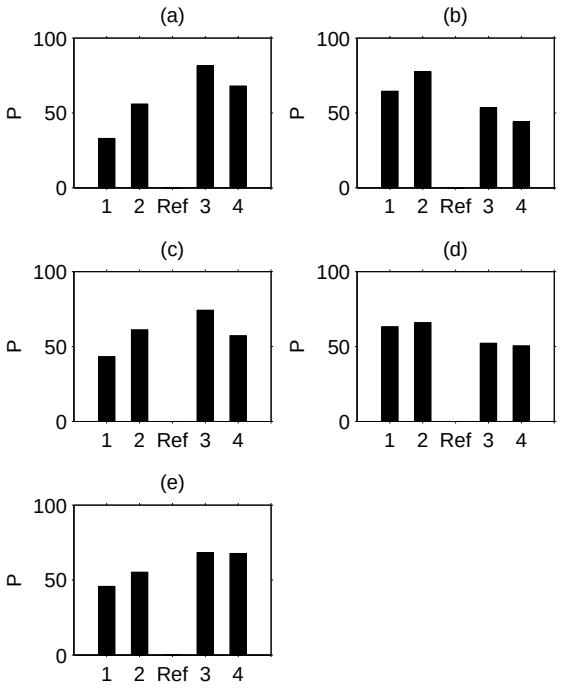

Figure 3: Distribution of the decisions for the frequency shift according to DEC 1 to 3, note: results containing crossed testings, for instance 1 vs. 4 against Ref
Identification of the reference vowels in test: (a)=[*i*], (b)=[*e*], (c)=[ɘ], (d)=[ɐ], (e)=[*o*]
For the Manipulation Index (abscissa) see Tab. 1

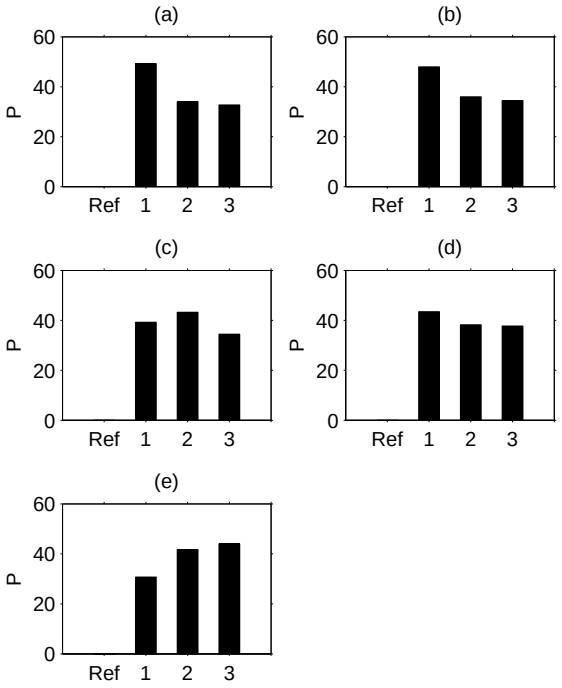

Figure 4: Distribution of the decisions for the bandwidth extensions according to DEC 1 to 3
Identification of the reference vowels in test: (a)=[*i*], (b)=[*e*], (c)=[ɘ], (d)=[ɐ], (e)=[*o*]
For the Manipulation Index (abscissa) see Tab. 1

especially on $F3$. Back and middle vowels are much more robust against $F3$ manipulations.
Results of this work suggest the separation into speaker dependent formants and vowel characteristic formants is correct only for back vowels. Front vowels, however, reveal logical structure of formants, so that lower and higher formants have to be *properly arranged* for a *correct* vowel quality perception.

REFERENCES

[1] G. Peterson, H. Barney, Control methods used in a study of the vowels, The Journal of the Acoustical Society of America 24 no 2 (1952) 175–184.

[2] P. Delattre, A. Liberman, F. Cooper, L. Gerstman, An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns, Word 8 (1952) 195–210.

[3] H. Kuwarabara, Y. Sagisaka, Acoustic characteristics of speaker individuality: Control and conversion, Speech Communication 16 (1995) 165–173.

[4] H. Pfitzinger, Towards functional modelling of relationship between the acoustics and perception of vowels, ZAS papers in Linguistics 40 (2005) 133–144.

[5] C. Stumpf, Die Struktur der Vokale, Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin (1918) 333–358.

[6] D. J. Broad, Piecewise-planar vowel formant distributions across speakers, J. Acoust. Soc. Amer. 69 no 5 (1981) 1423–1429.

[7] M. Barlow, F. Clermont (Eds.), A Parametric Model of Australien English Vowels in Formant Space, 8th Australian International Conference on Speech Science Technology, Australasian Speech Science and Technology ASSTA, Australia, Canberra, 2000.

[8] computer program "IPA Help", current version: 2.1 (2008).
URL `http://www.sil.org`

[9] J. D. Markel, A. H. Gray, Linear Prediction of Speech, Springer, 1976.

[10] H. K. Dunn, Methods of measuring vowel formant bandwidths, J. Acoust. Soc. Amer. 33, no 12 (1961) 1737–1746.

[11] P. G. Schlich, A method and SAS program for graphical representation of assessor performance, Journal of Sensory Science 9 (1994) 157Ú169.