Psycho-acoustical evaluation of pitch-marker positioning in natural speech

Institute for Perception Research
PO Box 513, 5600 MB Eindhoven

05.11.1996

Rapport no. 1132

Psycho-acoustical evaluation of pitch-marker positioning in natural speech

Mark M.J. Houben

Voor akkoord: Prof.dr. R. Collier

Psycho-acoustical evaluation of pitch-marker positioning in natural speech

A paper on a "korte stage" by:
Mark M.J. Houben
supervisor:
Reinier W. L. Kortekaas

November 5, 1996

Abstract

With the Pitch Synchronous OverLap and Add (PSOLA) technique, natural speech prosody can be manipulated. In this study, the influence of a relative shift ΔP of the 'analysis pitch-marker positions', used by PSOLA, has been investigated. Psychometric functions were measured by means of discrimination between speech signals synthesized using the original ($\Delta P = 0$) and the shifted ($\Delta P \neq 0$) analysis pitch-marker positions. These experiments are executed for lowered and raised pitch, both for the vowels /a/ and /i/. The results reveal that the pitch-markers can be shifted upto 15% of the distance between successive pitch-markers without the introduced distortions being aurally detectable.

The experimental results have been compared with an intensity-discrimination model, based on detecting intensity differences between excitation patterns. The used model was not adequate to fit the experimental data, even though it predicted the discrimination between synthetic single-formant signals well, in a comparable preceding experiment. It is hypothesized that besides differences in intensity there is at least one other cue on which subjects are able to discriminate between reference and signal.

Contents

Chapter 1

Introduction

The naturalness of text-to-speech systems, which are often based on concatenation of prerecorded natural diphones, can be increased by manipulation of prosody. This manipulation involves modification of the fundamental frequency (intonation) and duration (tempo and rhythm) of the speech signal. The ultimate target to strive for is to attain high synthesis quality and intelligibility and at the same time to reduce the computational cost. One technique that generally manages to achieve those competing aims quite well, is the Pitch Synchronous OverLap and Add (PSOLA) technique [Charpentier 90, Laroche 95]. Under PSOLA manipulation, phonemic content and voice quality is maintained even though the speech signal is manipulated with rather rough operations. However, sometimes annoying artefacts, such as hoarseness, accompany the alterations produced by the PSOLA manipulation.

In preceding psychophysical experiments in which *synthetic signals* were manipulated with the PSOLA technique [Kortekaas 96], the pitch-marker position was found to be not a very critical parameter in terms of detectability of distortions introduced by PSOLA. In practical applications, though, the position of the pitch-markers is generally considered to be a crucial factor for synthesis quality. The aim of the present study is to give a first impression about the relation between manipulation of *natural speech* and the psychophysical results.

Psychometric functions will be measured, in which sensitivity d' is presented as a function of the shift ΔP of the pitch-marker positions in relation to the signal maxima. The results will be compared with the synthetic signal manipulation experiments.

In the next chapter a brief explanation of the PSOLA technique will be given. Chapter 3 summarizes preceding findings of psychophysical experiments of interest for this research. Chapter 4 describes the methodology of the experiment and chapter 5 presents the resulting experimental data. A comparison of these results with a model based on detecting intensity differences between excitation patterns is the main topic of chapter 6.

Chapter 2

The PSOLA technique

The PSOLA technique consists of two phases:

(1) the analysis phase; the signal is decomposed into separate, but overlapping, segments,

(2) the synthesis phase; the segments are recombined by means of overlap adding.

In the analysis phase, the digitized speech waveform $x(k)$ is windowed at particular points of time (see figure 2.1A), resulting in a sequence of segments $x_n(k)$:

$$x_n(k) = h_n(k - pm_n)x(k)$$

where $h_n(k)$ is a sequence of pitch-synchronously positioned analysis windows (usually Hanning windows), centered around the successive points of time pm_n, called pitch-markers. These pitch-markers are either determined manually by inspection of the speech waveform or automatically by means of some local fundamental frequency ($F0$) estimation. The window lengths are usually set to be proportional to the local pitch period (temporal spacing between pitch-markers). In common PSOLA applications and in this study too, a factor 2 is chosen, which means that successive windows and thus successive segments have 50% overlap. The equation for a Hanning window (raised-cosine function) with an overlapping factor of 50% is:

$$h_n(k) = \frac{1}{2}(1 + \cos\frac{2\pi k}{2T_a}) \qquad k = -T_a \cdots T_a$$

where T_a is the analysis pitch-markers interval (local pitch period). In natural speech, T_a will not have a constant value due to variation in $F0$. The analysis windows will be asymmetrical and in the equation above, the factor $2T_a$ has to be replaced by the summation of the two consecutive pitch-markers intervals to the left and to the right, $T^-_{a,n} = pm_n - pm_{n-1}$ and $T^+_{a,n} = pm_{n+1} - pm_n$, respectively.

In the synthesis phase, segments are recombined after defining a new sequence of pitch-markers. A synthesized signal is produced by first assigning a decomposed segment to each of the new

pitch-markers (pitch-marker mapping) and then performing the sample-wise overlap-add operation, see figure 2.1C. The speech signal can be slowed down by repetition of segments and will be accelerated by eliminating segments. By changing the time intervals between pitch-markers, the fundamental frequency is changed. Increasing and decreasing the distance between pitch-markers lowers and raises $F0$, respectively.

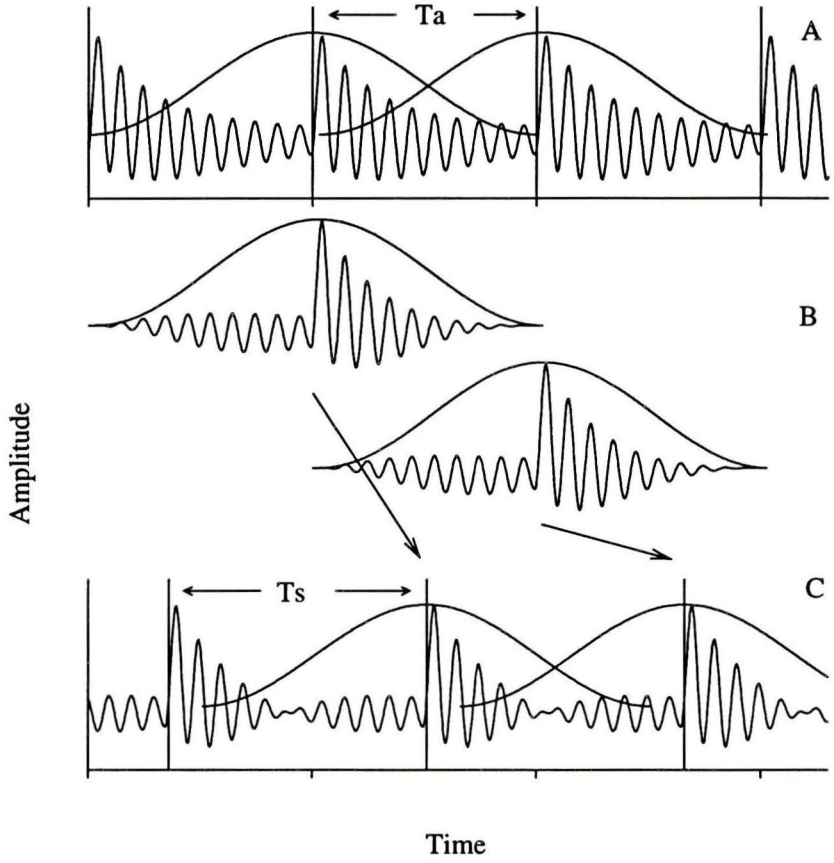

Figure 2.1: Illustration of the PSOLA technique, adapted from [Kortekaas 96]: Panel (A) shows the waveform of a signal. The thick vertical lines indicate the 'pitch synchronously' spaced analysis pitch-markers. The interval between two successive pitch-markers is indicated by T_a. The signal is decomposed into segments by windowing it with Hanning windows (cosine curves). In panel (B) two segments are shown. Panel (C) shows the recombination of these segments by means of overlap-adding at the new pitch-marker positions. The interval between two synthesis pitch-markers (thick vertical lines) is indicated by T_s.

In preceding psychophysical experiments (see [Kortekaas 96] and chapter 3), signals were generated by using a formant filter excited by a pulse train. The shift of the pitch-marker positions relative to the filter excitations was denoted by the parameter ΔP. The filter excitations coincided, to a first approximation, with the signal energy maxima. The analysis and synthesis window rates are defined by $F_{wa} = 1/T_a$ and $F_{ws} = 1/T_s$, respectively (analogous to the fundamental frequency $F0$), where T_a and T_s are the average distances in time between successive pitch-markers in the

analysis and synthesis phase, respectively. The relative change of window rate is defined as:

$$\Delta F = \frac{F_{ws} - F_{wa}}{F_{wa}} \times 100\%$$

Chapter 3

Preceding findings of psychophysical experiments

In [Kortekaas 96] some psychophysical experiments of interest for this research were executed. Psychometric functions with the analysis pitch-marker shift ΔP as experimental parameter were measured for synthetic single-formant signals for $\Delta F = -9.09\%$ and $\Delta F = +11.11\%$. These signals were generated by exciting a second-order digital resonator by a pulse train with an $F0$ of 100 Hz, equal to the analysis window rate F_{wa}. The formant frequency f_r was 1000 Hz with -3 dB bandwidth of 50 Hz. Some conclusions made in [Kortekaas 96] are summed up below.

- The thresholds for discrimination between unmanipulated synthetic signals and PSOLA manipulated synthetic signals with shifted analysis pitch-markers were found to be approximately $|\Delta P| = 25\%$. These thresholds are reasonably stable under level and formant frequency roving.
- The influence of a pitch-marker shift can be described well with an intensity discrimination model (to be explained in chapter 6). Although the differences between the single and multi-band model were small, the best results were obtained for the multi-band version, suggesting that discrimination was based on profile analysis.
- If intensity discrimination determines detectability, then thresholds for higher $F0$ values are expected to be lower. The results of informal tests for $F_{wa} = F0 = 250$ Hz confirmed this expectation: $|\Delta P|$ thresholds were between 10% and 20%, which is considerably smaller than 25%.

Chapter 4

Methodology

4.1 Stimuli synthesis

For the synthesis of the stimuli, a string of vowels (/a/ and /i/) sung by a male speaker (L. ten Bosch) over two octaves with a fairly constant pitch, were used. With the help of the software programme GIPOS on a Silicon Graphics Indigo workstation, a vowel /a/ and vowel /i/ with an average pitch of 161 Hz and 166 Hz, respectively, were cut out. High and low sung vowels sounded strained and some vowels (in the middle region too) sounded quavery. The chosen vowels, however, were sung in the normal register of the speaker and had a reasonably constant pitch.

Pitch marker locations were calculated by GIPOS using a method based on detecting local energy maxima by means of singular value decomposition [Ma 94]. In figure 4.1 the distribution of the instantaneous $F0$, which is the inverse of the distance in time between successive pitch-marker positions, is shown in a histogram. The $F0$ distribution for vowel /a/ is much broader than for vowel /i/. The variation around the average $F0$ is not systematic, i.e. the pitch fluctuates randomly and does not decrease or increase steadily from beginning to the end of the speech signal.

In the environment of SI (an interactive signal processing programme), the vowels were lowered and raised in pitch by means of the PSOLA technique discussed in chapter 2. These manipulations were performed using regularly spaced synthesis pitch-marker positions, calculated by inversion of the chosen synthesis pitch. As experimental parameter, the relative shift ΔP of the analysis pitch-marker positions with respect to the original pitch-marker distances calculated in GIPOS, was taken. In formula form:

$$\begin{aligned} \overline{pm}_n &= pm_n + \Delta P\,(pm_{n+1} - pm_n) \quad \text{if } \Delta P \text{ is positive} \\ \overline{pm}_n &= pm_n + \Delta P\,(pm_n - pm_{n-1}) \quad \text{if } \Delta P \text{ is negative} \end{aligned}$$

where $\overline{pm}_n$ is the position (in time) of the shifted n^{th} pitch-marker, pm_n is the position of the unshifted n^{th} pitch-marker and ΔP is the relative pitch-marker shift (in fraction or percentage). The division into two parts, one for positive and one for negative values of ΔP is neater than using only one formula. For example, when ΔP is +1 (or +100%), the shifted n^{th} pitch-marker

position $\overline{pm}_n$ equals the unshifted $(n+1)^{\text{th}}$ pitch-marker position pm_{n+1}. When ΔP is -1(or -100%), $\overline{pm}_n$ equals pm_{n-1}. If only one formula is used, the first one for example, the pitch-marker shift is correct for positive, but not for negative ΔP values. If instead of a relative pitch-marker shift, an absolute (constant) pitch marker shift of ΔP times the average pitch-marker distance is taken, the pitch-marker shift will be incorrect both for positive and negative values.

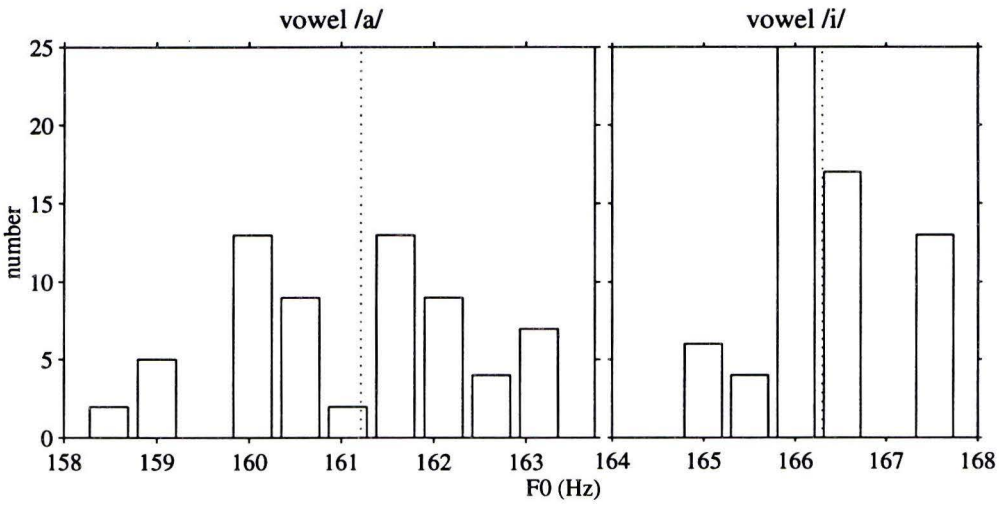

Figure 4.1: Histogram of instantaneous $F0$ (inverse of the distance in time between successive pitch-marker positions) for the vowels /a/ and /i/. The horizontal axis is continuous over the left (vowel /a/) and right panel (vowel /i/) and is subdivided in equal steps along the whole length. The bins are equally spaced. The $F0$ distribution for vowel /a/ is much broader than for vowel /i/. The dotted vertical lines represent the average $F0$ values.

The psychoacoustic measurements were set up in SISG (a script-based subsystem of SI). The sample frequency of the speech signals was doubled from 16 kHz to 32 kHz, used throughout the signal processing. The stimulus duration was 400 ms with ramping of the first and last 25 ms using a Hanning window. The separation between successive stimuli was 200 ms. After AD conversion, overall signal levels were adjusted to 70 dB SPL by means of analog attenuation. In order to increase stimulus uncertainty, level roving between intervals, uniformly distributed in the range of $\pm$ 5 dB, was applied. As mentioned in chapter 3, level roving did not 'dramatically' affect performances in preceding experiments [Kortekaas 96]. The purpose of level roving is to reduce the possibility of the subject fixating on a particular 'accidental' loudness difference between reference and signal.

Psychometric functions were measured by means of discrimination between speech signals synthesized using the original ($\Delta P = 0$) and shifted ($\Delta P \neq 0$) analysis pitch-marker positions, called 'reference' and 'signal', respectively. The pitch manipulation consisted of lowering and raising the pitch, both for the vowel /a/ and the vowel /i/. For vowel /a/, with an average analysis window rate F_{wa} of 161 Hz, the synthesis window rate F_{ws} for the lowered and raised condi-

tion was 127 Hz and 195 Hz, respectively. The relative change of window rate ΔF is -21.1% and $+21.1\%$ respectively. Vowel /i/, with an F_{wa} of 166 Hz, was lowered to 129 Hz and raised to 196 Hz, giving a ΔF of -22.3% and $+18.1\%$, respectively. By chosing these changes of window rate, the synthesized signals have a pitch corresponding with speech signals in the string of vowels (approximately two tone intervals lowered and one and a half raised). These vowels were used for measuring the discrimination between analysis pitch-marker shifted, PSOLA manipulated signals with regularly spaced synthesis pitch-markers (calculated) and original synthesis pitch-markers (taken from the corresponding vowels), see section 5.1.

4.2 Measurement procedure

During the experiments, the subjects were seated in a soundproof booth and received the stimuli over Beyer DT 990 headphones. They responded via an ordinary keyboard. Immediate feedback was given by revealing "correct" or "incorrect" after each trial and by reporting the percentage correctly answered trials per run. The method used for measuring the psychometric functions of discrimination between reference and signal, was the 3I3AFC odd-ball procedure (3 Intervals, 3 Alternatives, Forced Choice). The odd-ball interval contained the pitch-marker shifted signal. In each run the combination of ΔP, vowel type (/a/ or /i/) and pitch manipulation (lowered or raised pitch) was fixed. Each run consisted of 15 trials. The experiment with the whole set of runs, i.e. all tested combinations, was performed 4 times. Each condition thus was tested 60 times.

The experiments were performed by three subjects; MH (the author), RK (the supervisor) and JV. The first two subjects were familiar with the kind of stimuli used in this experiment. Subject JV additionaly performed one set of runs as learning phase, which was left out of the analysis.

Chapter 5

Experimental results

5.1 Preliminary experiments

Psychometric functions for discrimination between PSOLA-manipulated ('signal') and unmanipulated natural vowels ('reference') were measured as a baseline experiment. The experiment was performed by subjects MH and RK for both vowels and both directions of pitch manipulation. The analysis pitch-marker shift ΔP was varied between 0% and 50%. The synthesis pitch-marker positions were regularly spaced by calculating the positions with constant F_{ws}, or irregularly spaced by using the original pitch-marker positions of the corresponding vowel in the string of vowels. For both conditions a discrimination of 100% was observed, even for $\Delta P = 0$, indicating that the PSOLA technique introduces detectable distortions.

In another experiment, psychometric functions were obtained for discrimination between PSOLA-manipulated vowels using regularly spaced ('reference') and original synthesis pitch-markers ('signal'). The original pitch-markers were obtained by calculating the pitch-marker positions of the vowel, taken from the whole string of vowels, with a pitch corresponding to the pitch of the synthesized signal. The analysis pitch-marker shift ΔP was varied between 0% and 50% and was equal for reference and signal. The experiment was performed twice by subjects MH and RK for both vowels and for lowered, raised and fixed pitch. The fixed pitch actually corresponds to discrimination between the PSOLA-manipulated speech signal with regularly spaced synthesis pitch-marker positions with $F_{ws} = \text{average}(F_{wa})$ and the original speech signal. The results are shown in figure 5.1. The percentage correct responses Pc was converted to d', a measure of sensitivity of the subject for the physical difference between a reference and a signal. d' is defined within the theory of signal detection [Gelfand 90, Versfeld 92]. A table was used for conversion [MacMillan 91].

The psychometric functions of subject MH are very different from those of subject RK. Subject MH is able to discriminate between the reference (for which regularly spaced synthesis pitch-markers are used) and signal (for which original pitch-marker positions are used) for all conditions. Subject RK, however, is only able to discriminate for a few conditions (/a/ fixed, all ΔP values; /i/ lowered and /a/ raised, $\Delta P > 10\%$). No explanation has been found for this peculiar behaviour. Discrimination is probably based on detection of a non-static ('vibrating') pitch

in the signal. It is worth mentioning that both subjects are musically trained. It is likely that a not musically trained subject will find more difficulty in discriminating on the basis of the vibrating pitch cue. Vowel /a/, fixed pitch, seems to cause less discrimination difficulty than vowel /i/, fixed pitch. The distribution of $F0$ for vowel /a/ is much broader than for vowel /i/, as can be seen in figure 4.1. If $F0$ variation is a temporal cue on which subjects discriminate, it is likely to play a more dominant role in the vowel /a/ condition. This is in agreement with the experimental results.

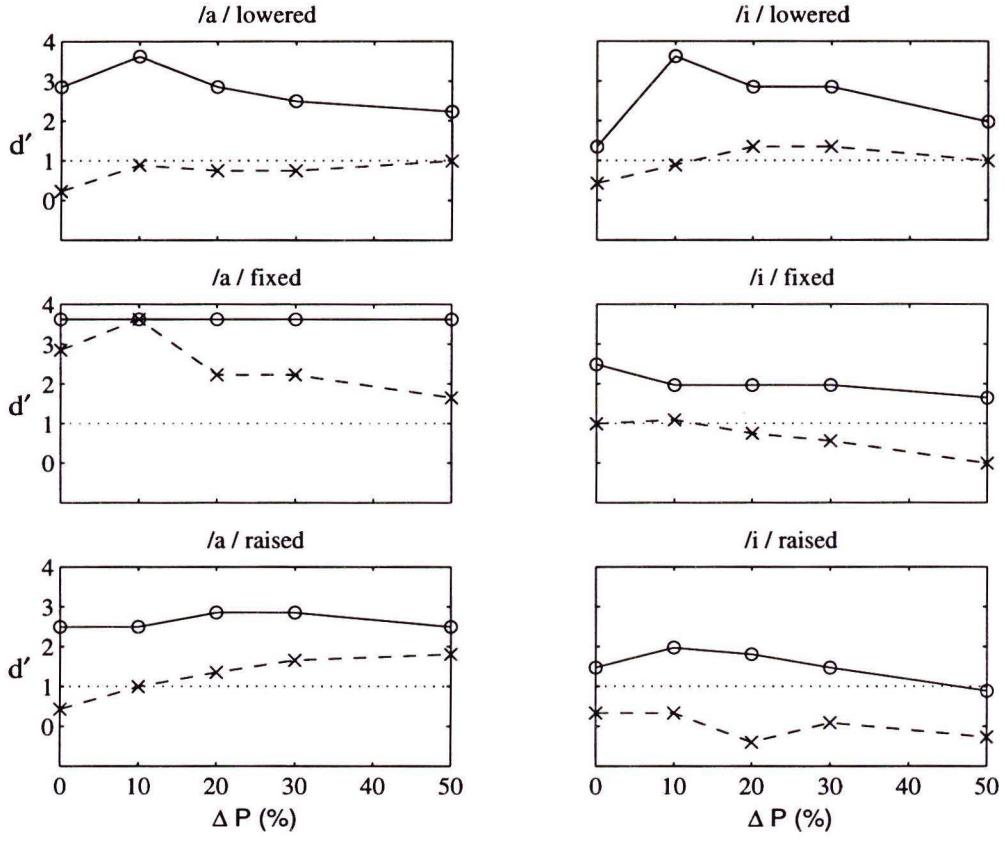

Figure 5.1: Psychometric functions for discrimination of PSOLA-manipulated speech signals with regularly spaced and original, irregularly spaced synthesis pitch-marker positions. Sensitivity d' is represented as a function of pitch-marker shift ΔP. The horizontal dotted line is the threshold $d' = 1$. Mean data of subject MH and RK are shown by circles and cross-signs, respectively. Left panels: vowel /a/, right panels: vowel /i/. Top panels: lowered pitch, middle panels: fixed pitch, bottom panels: raised pitch.

5.2 Pitch-marker shift experiments

Psychometric functions were obtained for discrimination of PSOLA-manipulated speech signals with unshifted and shifted analysis pitch-marker positions. The analysis pitch-marker shift ΔP was varied between -50% and +50%. The percentage correct responses Pc was converted to d'. The standard deviation of Pc over the 4 sets of runs was calculated and converted to d' too. The

mean d' values and standard deviations, as a function of the pitch-marker shift ΔP are represented, per subject, in the figures below (figures 5.2, 5.3 and 5.4). Figure 5.5 shows the results for all subjects combined.

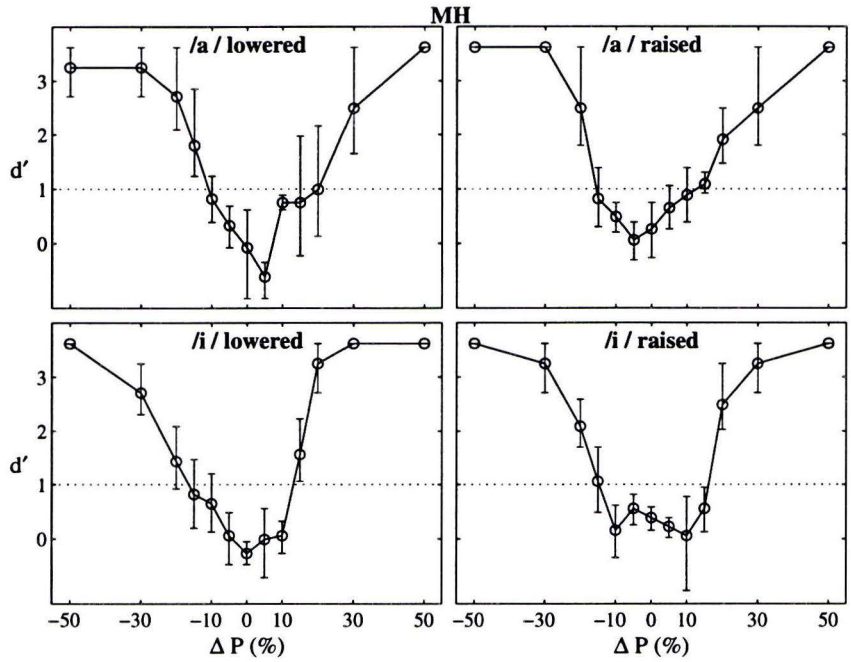

Figure 5.2: Psychometric functions for discrimination of PSOLA-manipulated speech signals with unshifted and shifted analysis pitch-marker positions, for subject MH. Sensitivity d' is represented as a function of pitch-marker shift ΔP. The horizontal dotted line is the threshold $d' = 1$. Mean data are shown by circles, standard deviations by vertical bars. Top panels show the results for vowel /a/, bottom panels for vowel /i/. Left panels show the results for lowered pitch, right panels for raised pitch.

The discrimination thresholds, estimated by taking the ΔP values at the points of intersection of the horizontal line $d' = 1$ with the curves of the psychometric functions, are listed in the table below.

	/a/ lowered		/a/ raised		/i/ lowered		/i/ raised	
MH	-11	+19	-16	+13	-17	+13	-15	+16
RK	-13	+15	-17	+21	-23	+10	-13	+15
JV	-15	+15	-18	+21	-22	+12	-16	+23
average	-13	+16	-16	+18	-21	+12	-15	+17

Table 5.1: Estimated discrimination thresholds for the pitch-marker shift experiments.

The differences across the subjects are small, with the exception of the raised /a/ for which the sensitivity of subject MH in the $\Delta P > 0\%$ region is higher and the raised /i/ for which the sensitivity of subject JV in the $\Delta P > 0\%$ region is smaller than the other two subjects. The psychometric

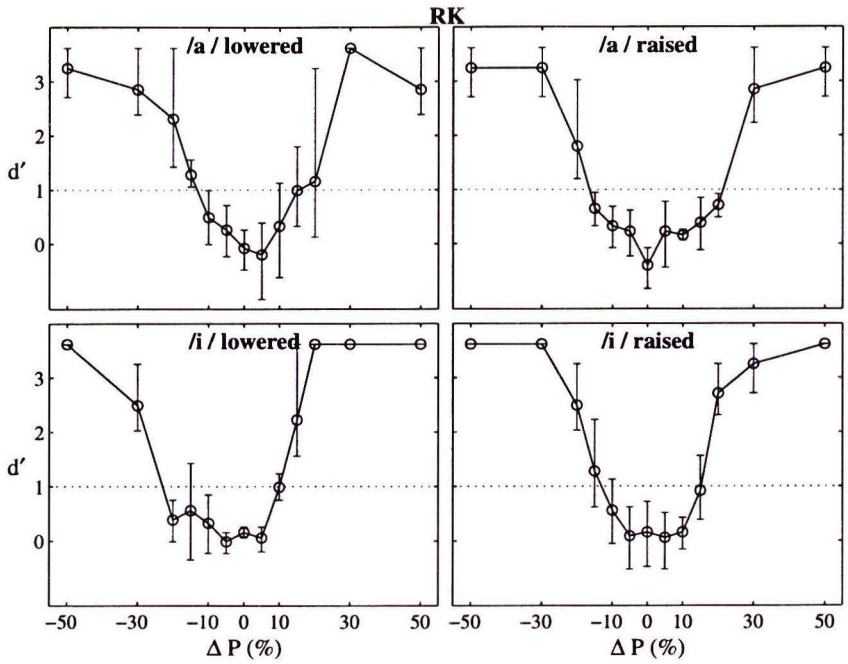

Figure 5.3: Psychometric functions for discrimination of PSOLA-manipulated speech signals with unshifted and shifted analysis pitch-marker positions, as in figure 5.2 but for subject RK.

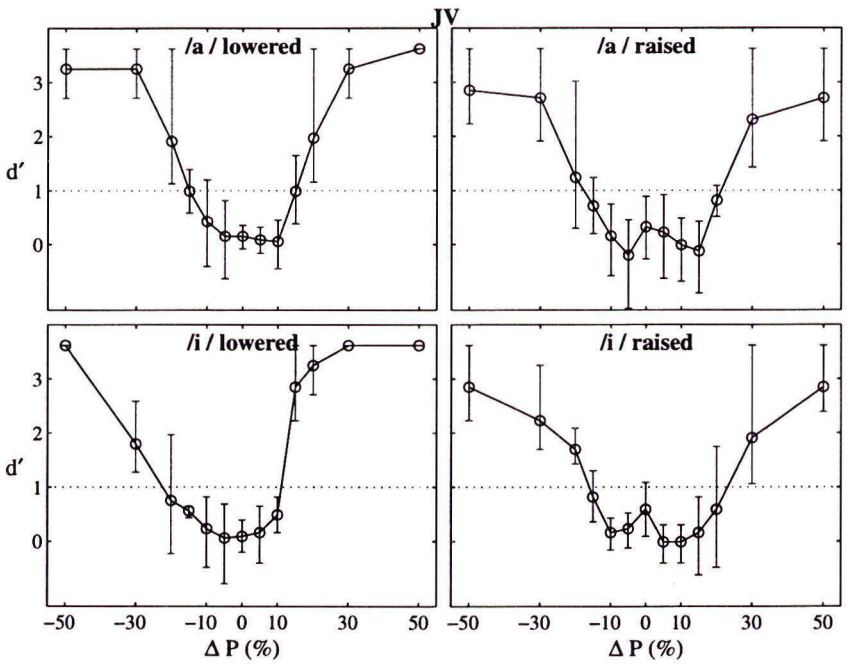

Figure 5.4: Psychometric functions for discrimination of PSOLA-manipulated speech signals with unshifted and shifted analysis pitch-marker positions, as in figures 5.2 and 5.3 but for subject JV.

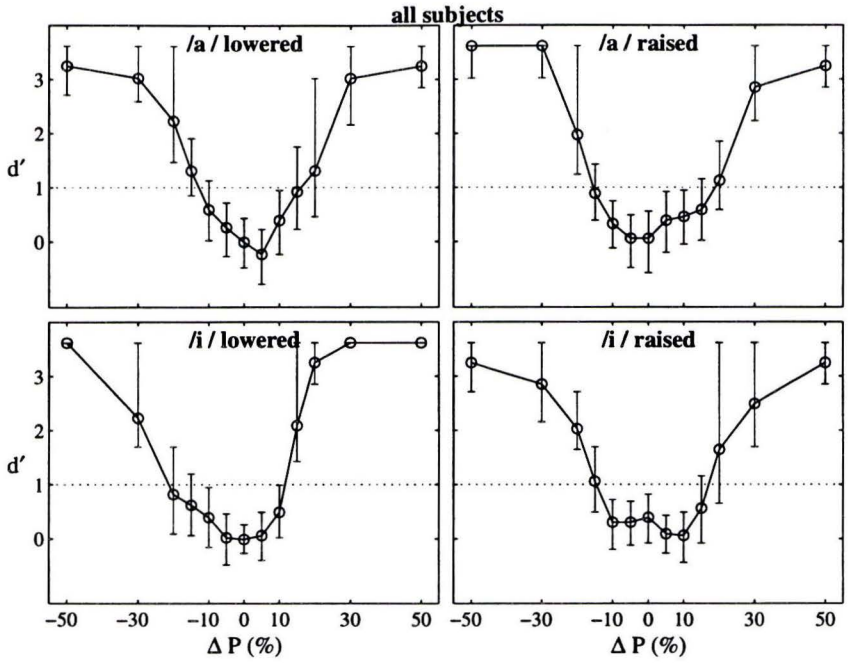

Figure 5.5: Psychometric functions for discrimination of PSOLA-manipulated speech signals with unshifted and shifted analysis pitch-marker positions, for all three subjects, shown in figures 5.2, 5.3 and 5.4, combined.

functions are, for the most part, symmetric around $\Delta P = 0\%$, except for the lowered /i/ curve which is skewed to the left (negative ΔP direction).

5.3 Discussion

The discrimination performance approximately reaches threshold at an absolute value of ΔP of 15%. The position of the pitch-markers thus is not a very critical parameter, nevertheless it should be placed near the signal energy maximum, at least within the interval of -15% and +15% of the distance between the two successive pitch-marker positions. This interval is smaller than the threshold of $|\Delta P| = 25\%$ for synthetic single-formant signal discrimination with $F0 = 100$ Hz, but comparable with the $|\Delta P|$ threshold values between 10% and 20% for synthetic signals with $F0 = 250$ Hz, found by [Kortekaas 96] (see chapter 3). By manipulating natural vowels (with additional information, more formants for example) instead of synthetic single-formant signals, the distortions introduced by the PSOLA manipulation may increase the discrimination performance. But it is also possible that the lower thresholds of the natural vowel experiments relative to the $F0 = 100$ Hz synthetic signal experiments are caused by a higher $F0$.

Chapter 6

Model

The main question in this chapter is how the discrimination results relate to the results obtained with a psycho-acoustical model. The investigated model is an intensity-discrimination model based on detecting intensity differences between excitation patterns [Florentine 81]. This model was able to describe the experimental results of the variation of the pitch-marker positions, using synthetic single-formant signals, pretty well, see [Kortekaas 96] and chapter 3.

6.1 Theory

The intensity-discrimination model only takes spectral cues into account. For both the reference and the signal, excitation patterns are calculated by means of filtering using a Gammatone filter bank [Patterson 87]. The level differences between the excitation patterns $\Delta L_{E,i}$ per channel i, are determined. These channel bandwidths correspond with the auditory critical bands. The partial sensitivity in channel i is d'_i. The model assumes that d'_i is proportional to $\Delta L_{E,i}$ with a constant factor k which is the same for all channels. The overall sensitivity d' is derived from the partial sensitivities d'_i, $i = [1, \ldots, N]$, where N is the number of channels. Two different versions are investigated: a single-band and a multi-band version. In the single-band version the overall sensitivity d' is equal to the maximum of the partial sensitivities:

$$d' = \max_{i=1,\ldots,N}(d'_i) = k \cdot \max_{i=1,\ldots,N}(\Delta L_{E,i}) = k \cdot D_{max}$$

In the multi-band version, partial sensitivities are combined according to [Durlach 86]:

$$d' = (\sum_{i=1}^{N} {d'_i}^2)^{\frac{1}{2}} = k \cdot (\sum_{i=1}^{N} {\Delta L_{E,i}}^2)^{\frac{1}{2}} = k \cdot D_{sum}$$

According to the two formulas, d' is linearly related to D_{max} or D_{sum}. By performing a linear regression on the experimental data (expressed in d') in dependence on D_{max} or D_{sum}, the predictive power of the model can be investigated. The slopes k of the linear regression equations are calculated, per subject, in [Kortekaas 96], for a comparable pitch-marker shift experiment, with the main difference that synthetic single-formant signals, instead of natural speech vowels, were

used. The average over all three subjects is $k = 0.17$ for the multi-band model and $k = 0.28$ for the single-band model.

6.2 Results

Excitation patterns were calculated for the references (unshifted) and signals (shifted pitch-markers) used in the experiment. D_{sum} and D_{max} for the multi-band and single-band version, respectively, were calculated out of the level differences between the excitation patterns of reference and signal. The d' predictions of the model, calculated by multiplying k by D_{sum} or D_{max}, are represented in figure 6.1. Shown are both the multi channel variant (solid line, o markers) and the single, maximum channel variant (dashed line, + markers), as well as the in figure 5.5 represented measured psychometric functions averaged over all subjects (dot-dashed line, x markers).

The predicted curves are more flat and shallow than the psychometric functions measured in the experiments. In the top left panel (/a/ lowered), an unexplained peak value is present with d' values for the multi-band and single-band model of 5.3 and 4.5, respectively.

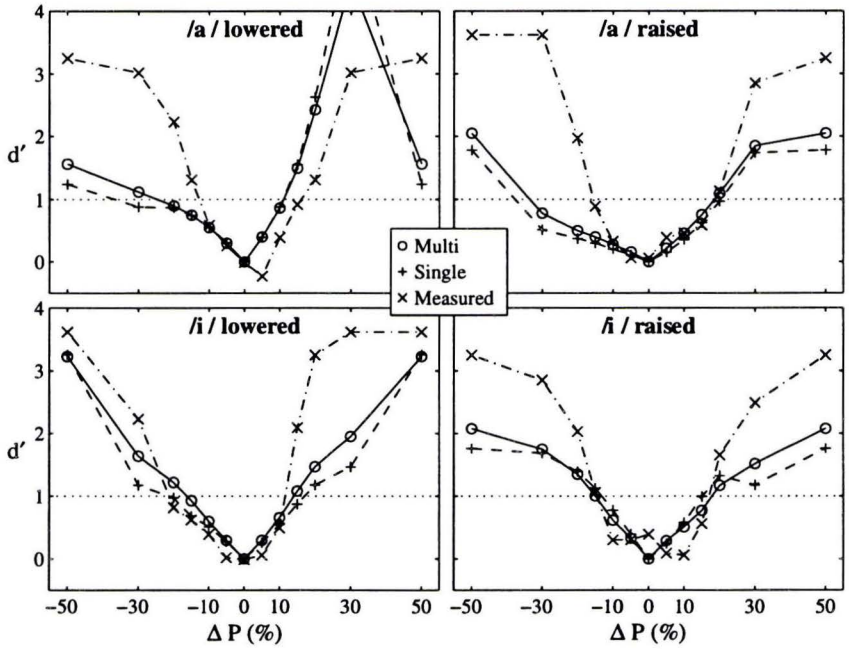

Figure 6.1: Psychometric functions for discrimination of PSOLA-manipulated speech signals with unshifted and shifted analysis pitch-marker positions, predicted by the intensity-discrimination model. Sensitivity d' is represented as a function of pitch-marker shift ΔP. The horizontal dotted line is the threshold $d' = 1$. Shown are the multi channel variant (solid line, o markers), the single, maximum channel variant (dashed line, + markers) and the measured psychometric functions of the experiments (averaged over all three subjects). Top panels show the results for vowel /a/, bottom panels for vowel /i/. Left panels show the results for lowered pitch, right panels for raised pitch.

The estimated threshold values (points of intersection of the horizontal line $d' = 1$ with the prediction curves) are listed in table 6.1.

	/a/ lowered		/a/ raised		/i/ lowered		/i/ raised	
Multi-band model	-25	+11	-34	+18	-15	+13	-15	+17
Single-band model	-37	+11	-38	+20	-21	+17	-13	+15
Experiment average	-13	+16	-16	+18	-21	+12	-15	+17

Table 6.1: Thresholds predicted by the intensity-discrimination model for both the multi-band and the single-band variant and average thresholds measured by the experiment.

6.3 Discussion

The differences between the predicted d' values of the single and multi-band model are small. Taking this small difference into account, the multi-band model resembles the measured psychometric functions slightly better. Comparison of the average measured and the model predicted thresholds in table 6.1 reveals that the thresholds for vowel /i/ are estimated well by both models but for vowel /a/ the model predictions deviate greatly from the measured thresholds; the model threshold values are higher. The psychometric functions generated by the model are less steep at the sides.

One possible hypothesis is that the slope k of the linear regression equation is too small. If k increases, d' will increase too, resulting in a better fit of the model for the vowel /a/. But this will inevitable result in increased d', and thus decreased thresholds for the vowel /i/, while the predicted psychometric functions resembled the measured psychometric functions reasonably well with usage of the previous k for vowel /i/. So this hypothesis is not likely to be true.

Another possibility is that k is not a constant factor. If k is not equal for both vowels, the predicted psychometric functions for vowel /i/ can remain unchanged while those of vowel /a/ can be adapted. But k is a measure of sensitivity of the ear for intensity level differences within a channel (bandwidth). It would be very odd if the auditory sensitivity changes when the level differences are caused by presenting a vowel /a/ instead of a vowel /i/. The difference between the prediction performance of the model for vowel /a/ and /i/ may be caused by spectral differences such as position of the formants. If instead of the assumed equalness of k for all channels, k varies from channel to channel (but is equal for both vowels), the model may be able to fit the psychometric functions of both vowels. The disadvantage, however, is that a k dependend on the spectral position of the channel results in many unknown parameters. With these parameters it is not difficult to attain a better fit, but it is to be doubted if the resulting model is still a usefull representation of the auditory perception.

A third hypothesis is that the used filter bandwidths are too broad. Narrowing the bandwidths may result in an increase of detectability of intensity differences by the model. Due to a higher resolution, a notch in the amplitude spectrum, introduced by the PSOLA manipulation, will easier be detected. But this change of the width of the channels undermines the underlying ideas of the model that the channels correspond with the critical auditory bands.

All this implies that the used model is not adequate to fit the experimental data, i.e. besides differences in intensity there is at least one other cue (a temporal cue, for example), on which subjects are able to discriminate between reference and signal.

Although this model is adequate for describing the results of the pitch-marker shift experiments using synthetic single-formant signals, this is obviously not the case with natural vowels. The used intensity-discrimination model only takes spectral cues into account. Implementing a model which focusses on temporal cues will give more information about the discrimination behaviour and is therefore a recommendation for further research.

The duration of the used signals was 400 ms. If shorter signals will be used (which may occur in practical applications), the possible additional temporal cue may be less eminently present.

Chapter 7

Conclusions

- The PSOLA technique introduces detectable distortions, even for $\Delta P = 0$.
- The measured discrimination thresholds for the pitch-marker shift experiment are approximately $|\Delta P| = 15\%$. The position of the pitch-markers is not a very critical parameter for synthesis quality.
- The thresholds for vowel /i/ are estimated reasonably well by an intensity discrimination model but for vowel /a/ the model predictions deviate greatly from the measured thresholds. The multi-band version of the model resembles the measured psychometric functions better than the single-band version, but differences are very small.

  Although this model is adequate for describing the results of the pitch-marker shift experiments using synthetic single-formant signals, this is not the case with natural vowels, i.e. besides differences in intensity there is at least one other cue (temporal cues, for example), on which subjects are able to discriminate between reference and signal.
- The used intensity-discrimination model only takes spectral cues into account. Implementing a model which focusses on temporal cues will give more information about the discrimination behaviour and is therefore a recommendation for further research.
- Research on the influence of the spectral content of a natural speech signal (position of the formants, for example) on the discrimination behaviour of subjects, will give further information about the perceptual effects of PSOLA manipulation.

Bibliography

[Charpentier 90] F. Charpentier, E. Moulines,
"Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones." *Speech Commun.* **9**, 453-467, (1990)

[Durlach 86] N.I. Durlach, L.D. Braida, Y. Ito,
"Towards a model for discrimination of broadband sounds." *J. Acoust. Soc. Am.* **80**, 63-72, (1986)

[Fassel 93] R. Fassel,
"SI, a pocket calculator for signals: introduction." Handleiding no. 125, Institute for Perception Research, Eindhoven, The Netherlands, (1993)

[Fassel 96] R. Fassel, S. van de Par, M. van der Heijden,
"SISG, a psychoacoustic measurement system." Handleiding no. 139, Institute for Perception Research, Eindhoven, The Netherlands, (1996)

[Florentine 81] M. Florentine, S. Buus,
"An excitation-pattern model for intensity discrimination." *J. Acoust. Soc. Am.* **70**, 1646-1654, (1981)

[Gelfand 90] S.A. Gelfand,
Hearing: an introduction to psychological and physiological acoustics. 2nd rev. and enl. ed., Dekker, INC., New York, (1990)

[Kortekaas 96] R.W.L. Kortekaas, A. Kohlrausch,
"Psychoacoustical evaluation of the PSOLA speech-waveform manipulation technique using single-formant stimuli." Submitted to *J. Acoust. Soc. Am.*, (1996)

[Laroche 95] J. Laroche, E. Moulines,
"Non-parametric techniques for pitch-scale and time-scale modification of speech." *Speech Commun.* **16**, 175-205, (1995)

[Ma 94] C. Ma, Y. Kamp, L.F. Willems,
"A frobenius norm approach to glottal closure detection from the speech signal." *IEEE Trans. Speech Audio Proc.* **2**, 258-265, (1994)

[MacMillan 91] N.A. MacMillan, C.D. Creelman,
Detection theory: a users guide. Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney, (1991)

[Patterson 87] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice,
"An efficient auditory filterbank based on the gammatone function." in *Appendix B of SVOS Final Report: The auditory filterbank*, volume APU report 2341, (1987)

[Versfeld 92] N.J. Versfeld,
"On the auditory discrimination of spectral shape." Unpublished PhD thesis, Institute for Perception Research, Eindhoven, The Netherlands, (1992)