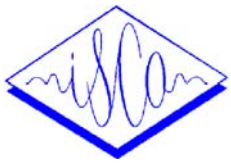

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

PSYCHOPHYSICAL EVALUATION OF PSOLA: NATURAL VERSUS SYNTHETIC SPEECH

R. Kortekaas and A. Kohlrausch

IPO Center for Research on User-System Interaction
P.O. Box 513 - 5600 MB Eindhoven - The Netherlands
E-mail: kortekaa@ipo.tue.nl kohlraus@ipo.tue.nl

ABSTRACT

This paper presents the results of psychophysical experiments dealing with pitch-marker positioning within the Pitch Synchronous OverLap and Add (PSOLA) framework. Sustained natural vowels were PSOLA-modified in fundamental frequency. The experiments were aimed at determining the auditory sensitivity to (1) deterministic shifts of either all or single pitch markers within a sequence, and (2) random shifts of all pitch markers ("jitter"). As for deterministic shifts of all pitch markers, the results were in reasonable agreement with results obtained previously for synthetic formant signals. For deterministic shifts of single pitch markers, thresholds depended on position in the sequence. Detection thresholds for jittered shifts were comparable to thresholds for detecting jitter in pulse trains. The ranking of the thresholds for these three conditions indicated that the auditory system is more sensitive to dynamic (modulation) cues rather than to static (timbral) cues arising from shifts in pitch-marker positioning.

1. INTRODUCTION

The Pitch-Synchronous OverLap and Add (PSOLA) technique is a well-known method for time-scale and pitch-scale modification of natural speech [1]. The present paper focuses on the time-domain implementation (TD-PSOLA) which will be referred to as PSOLA in the following. Even though PSOLA has found widespread application, little is known about the perceptual consequences of the PSOLA operations such as pitch-marker positioning.

PSOLA modification of an input speech signal is based on determining pitch markers, which indicate boundaries of local pitch periods in the case of voiced speech. Pitch-marker positioning is commonly assumed to be an important factor for synthesis quality. The present study was aimed at psychophysically determining the sensitivity of the human auditory system to shifts in pitch-marker position. The findings of these experiments provide information about the required accuracy of pitch-marker determination.

The results reported on here are an extension to those presented in [2] where, among other things, the role of pitch-marker shifts in synthetic single-formant signals [3] was studied . Pitch-marker shifts were defined as shifts relative to the pulses that excited the formant filter. All pitch markers were shifted equally. Detection thresholds for pitch-marker shifts were about 25 % of the fundamental period, for a fundamental frequency (*F0*) of 100 Hz. For an *F0* of 250 Hz thresholds were (informally) measured to be approximately 10 %. Discrimination performance was also seen to increase monotonically with increasing pitch-marker shift. This performance could be described well by a psychoacoustic model based on excitation pattern differences [4].

In experiment 1 of the present paper, similar experiments were performed using natural sustained vowels [5]. In contrast to the synthetic signals, such signals fluctuate (slightly) in *F0*, formant frequencies and level over time. The aim was to determine whether this non-stationarity also resulted in monotonicity and comparable thresholds of discrimination. In experiment 2 discrimination thresholds for shifting single pitch makers were measured. In contrast to shifting all pitch markers, such a shift yields auditory cues that dynamically vary over the duration of the signal. In experiment 3 thresholds were measured for detecting random shifts ("jitter") imposed on the pitch-marker sequence. Such shifts can be conceived of as small errors in the (local) *F0* estimate and also introduce dynamic cues.

2. GENERAL METHODS

2.1 Pitch Markers

The pitch markers of the natural sustained vowels were determined by (1) estimating the local *F0* and (2) determining the local energy maxima [6]. For each signal this resulted in a sequence of pitch markers P_i^a (i = 1,N), where superscript "a" indicates "analysis" and N is the number of pitch markers in the sequence. In experiment 1 each pitch marker P_i^a was shifted over a relative amount ΔP given by:

$$\overline{P}_i^a = \begin{cases} P_i^a + \Delta P\left(P_{i+1}^a - P_i^a\right) & \text{if } \Delta P \geq 0 \\ P_i^a - \Delta P\left(P_i^a - P_{i-1}^a\right) & \text{otherwise} \end{cases}$$

ΔP thus expresses a fraction of the local fundamental period and will be presented as percentage in the following. In experiment 2 just a single pitch marker was shifted according to the formula given above.

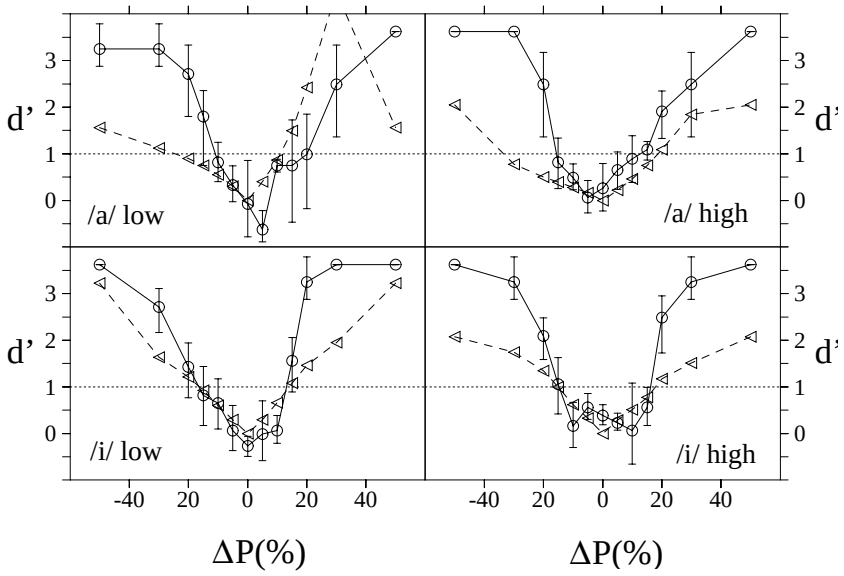

Figure 1: Psychometric functions for subject MH (circles) showing discrimination sensitivity as a function of ΔP. Vertical bars indicate standard deviations. Model predictions are shown by the triangles.

In experiment 3 pitch markers were randomly shifted according to :

$$\overline{P}_i^a = P_i^a + \mathrm{J} \cdot T0$$

where *T0* is the mean fundamental period. **J** is a spectrally white, Gaussian random variable with zero mean and standard deviation σ. This standard deviation will be presented as a percentage of *T0*.

2.2 Original Signals

A (non-professional) male speaker uttered a sequence of isolated vowels /a/ and /i/ at semitone intervals over one octave. Recordings were made in a low-reverberant, quiet room using B&K microphones and a DAT recorder. The signals were stored on computer hard disk after low-pass filtering at 8 kHz (sample rate 48 kHz). One realization of both the vowel /a/ and /i/ from the sequence was chosen that fell within the normal register of the speaker. Taking the pitch-marker intervals as measures of the (instantaneous) *F0*, the vowel /a/ had an average of 161 Hz with a range 158 to 163 Hz. The vowel /i/ was slightly more stable: average *F0* 166 Hz and a range of 165 to 167 Hz.

2.3 Experimental Stimuli

Using PSOLA, two modified vowels /a/ were synthesized having an *F0* of 127 and 195 Hz. These stimuli will be referred to as /a/-low and /a/-high, respectively. Likewise, the *F0* of the vowel /i/ was modified to 129 and 196 Hz (/i/-low and /i/-high). These *F0* shifts amount to approximately 3 semitones. The vowels were synthesized with a constant *F0*, i.e., the synthesis pitch-marker sequence P_i^s had constant intervals.

Stimuli were 400 ms in duration where the first and last 25 ms were ramped using a raised-cosine window. The ramp duration in experiment 2 was 15 ms. The stimuli were presented to the subjects over Beyer DT990 headphones with a mean overall presentation level of 70 dB SPL. On each presentation the level was roved within +/- 5 dB in order to rule out the use of possible loudness cues. Subjects were seated in a sound-proof booth and received immediate feedback after each trial.

2.4 Measurement Procedures

In all experiments PSOLA-modified vowels using a shifted or jittered pitch-marker sequence ("test") had to be discriminated from vowels modified using the original sequence ("reference"). A 3I-3AFC paradigm was used in which one test and two reference stimuli were presented to the subject in random order. The subject's task was to indicate which interval contained the deviant "test" stimulus.

In experiment 1 psychometric functions were measured in which discriminability was determined as a function of Δ*P*. For each condition, all subjects performed at least one set of measurements, containing 15 trials, as a practice. The data presented below are the means (and standard deviations) over the final four sets of measurements. Instead of presenting percentage correct *Pc*, data will be presented in terms of the discrimination index *d'*. The discrimination threshold corresponds to *d'* = 1.

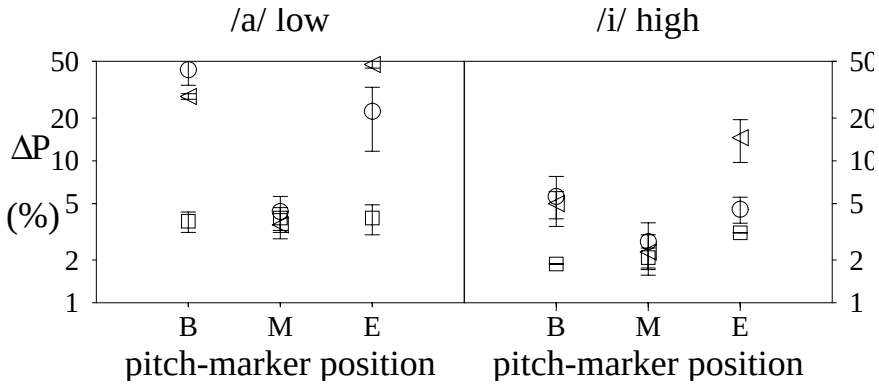

Figure 2: ΔP thresholds for shifting single pitch markers as a function of pitch-marker position (see text for details). Data for subject HH are shown by circles, for PK by triangles, and RK by squares. Vertical bars indicate standard deviations.

In experiment 2 and 3 an adaptive 3I-3AFC paradigm was used to measure thresholds for ΔP (single shifts) and σ, respectively. The level of ΔP or σ was decreased after two correct responses, and raised after one incorrect response ("two-up, one-down"). The minimal step size for ΔP and σ was 1.25 % and 0.1 %, respectively. Note that with a sampling rate of 48 kHz and a (mean) fundamental period of about 6 ms, shifting a pitch marker over one sample corresponds to $\Delta P \sim 0.3$ %. All subjects performed at least one threshold measurement for each condition as a practice. The data presented below are the meansand standard deviations of the last three measurements.

3. UNIFORM SHIFTS

3.1 Results

Three subjects participated in this experiment. Subjects mostly reported timbral cues ("nasality") as their discrimination criterion. Because the data were sufficiently similar among subjects, only the psychometric functions for subject MH are shown in Figure 1. As for the synthetic single-formant signals [2], the psychometric functions show monotonicity as a function of ΔP. The shape of the functions does not reveal a systematic difference between raising or lowering *F0*, or between /a/ en /i/.

The detection thresholds (i.e., the value of ΔP for which $d' = 1$) amount to about 15 %. In [2] detection thresholds were reported of approximately +/- 25 % for an *F0* of 100 Hz. Thresholds for higher *F0* values were expected to be lower. Because the *F0* of the natural vowels is about 160 Hz, this finding is thus in agreement with the previous results.

3.2 Modeling

The psychometric functions for reported in [2] could be described well by using an intensity-discrimination model [4]. Such a model calculates the difference between the excitation patterns of the "test" and "reference" stimulus. The excitation pattern is derived by analyzing the stimulus by means of an (auditory) filterbank and calculating the power within each channel. In this way the model outcome only depends on the power spectrum of the stimulus.

Figure 1 also shows the psychometric function predicted by a multiband model in which all channels of the filterbank are taken into account. This model was gauged to the previously reported psychometric functions for synthetic stimuli. Except for negative shifts in the case of raising the vowel /a/, all thresholds are predicted rather accurately. With just a few exceptions, however, the predicted psychometric functions for values above the threshold lie substantially below the measured *d'* values.

4. SINGLE SHIFTS

Threshold measurements were performed for single shifts of the fourth, $(N/2)^{th}$, and $(N-4)^{th}$ pitch marker (denoted by "B", "M", and "E", respectively). These conditions were investigated for the /a/-low and /i/-high condition.

Three subjects participated in the experiment and the results are shown in Figure 2. The cue they reported mostly was a (rough) discontinuity in the "test" stimulus. The thresholds for shifting the middle pitch marker ("M") are comparable for the three subjects. These thresholds are about 2 to 5 % which is a factor 3 lower than for uniform shifts. This suggests that the auditory system is more sensitive to the dynamic changes introduced by single shifts than to the (almost) static cues introduced by uniform shifts. The threshold for the /a/-low condition seems to be higher than for the /i/-high condition.

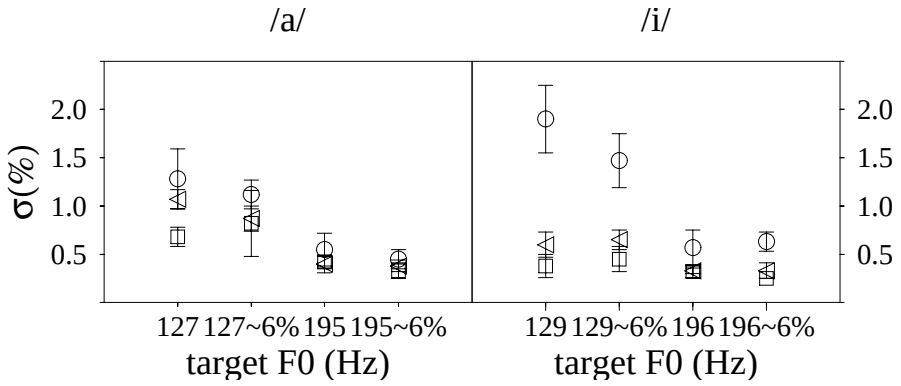

Figure 3: Detection thresholds for σ as a function of target F0. Roving of target F0 is indicated as in "127~6 %" (see text for details). Symbols are used as in Figure 2

Thresholds for the "B" and "E" conditions are generally higher than for "M". The variability among subjects is, however, rather high, especially for the /a/-low condition. This finding is in agreement with data presented in [7]. In that study *pulse trains* were used as stimuli and thresholds were measured for (random) changes in the inter-pulse intervals. In the case of shifting single pulses, higher thresholds were reported when shifting leading and trailing pulses rather than central pulses (see figure 5 in [7]).

5. JITTERED PM SEQUENCES

The participating subjects were the same as in experiment 2. They reported roughness (for moderate and large σ) and unsteadiness (near threshold) as discrimination cues. The σ thresholds shown in Figure 3 for jitter RMS are about 0.5 to 1 %. The lowest thresholds are observed for the /a/-high and /i/-high conditions. These two findings are in agreement with the data on jitter detection of (filtered) pulse trains presented in [7]. The Δ*P* thresholds for single "M" shifts are about 4 times σ at threshold. Therefore, discrimination at near-threshold levels may have been based on detecting single shifts.

Figure 3 additionally shows thresholds measured with roving of the synthesized ("target") *F0* over +/- one semitone (6 %). The resulting thresholds do not differ markedly from the constant-*F0* condition which indicates that subjects did not base their discrimination on differences in pitch.

6. CONCLUSIONS

The discrimination performance for uniform pitch-marker shifts in sustained natural vowels is qualitatively similar to the performance for synthetic formant signals reported previously. The performance could at least be partly explained by model predictions. The auditory sensitivity for single shifts is found to be position dependent. Central shifts are most easily detectable and thresholds are about three times lower than for uniform shifts. The thresholds for random (jittered) shifts are lowest. Jittered shifts can provoke clear sensations of roughness even at rather low levels of the jitter RMS.

7. ACKNOWLEDGEMENTS

The authors would like to thank Mark Houben for carrying out part of the experiments. The comments of Andrew Oxenham were greatly appreciated.

8. REFERENCES

[1] E. Moulines and J. Laroche, "Non-parametric tehniques for pitch-scale and time-scale modification of speech", *Speech Communication*, Vol. 16, pp. 175-205, 1995.

[2] R. Kortekaas and A. Kohlrausch, "Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli", *Journal of the Acoustical Society of America*, Vol. 101, pp. 2202-2213, 1997.

[3] D. Klatt, "Software for a cascade/parallel formant synthesizer", *Journal of the Acoustical Society of America*, Vol. 67, pp. 971-995, 1980.

[4] J. Gagne and P. Zurek, "Resonance frequency discrimination", *Journal of the Acoustical Society of America*, Vol. 83, pp. 2293-2299, 1988.

[5] M. Houben, "*Psycho-acoustical evaluation of pitch-marker positioning in natural speech*", internal IPO report 1132, Eindhoven, 1996.

[6] C. Ma, Y. Kamp and L. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal", *IEEE Transactions on Speech and Audio Processing*, Vol 2, pp. 258-265, 1994.

[7] B. Cardozo and R. Ritsma, "On the perception of imperfect periodicity", *IEEE Transactions on Audio and Electroacoustics*, Vol 16, pp. 159-164, 1968.