Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 20 (1996), 139-142.

SUMMARY OF A PH.D. THESIS DEFENDED IN 1996

Pitch perception in speech: a time domain approach

author: Henning Reetz [1]
promotores: L.C.W. Pols and A. Lahiri [1]
date of defence: 13 november 1996

Summary

This thesis deals with the perception of pitch of speech signals and ways to measure this pitch. The term 'pitch' is used as a collective term for speech production (where it is used for the fundamental frequency of quasi-periodically vibrating vocal folds), for acoustic transmission (where it is used for the periodicity of signals), and for speech perception (where it is used for the perceived pitch of speech signals). In the last case, the frequency of the perceived pitch of a signal is expressed by the fundamental frequency of a reference signal with a rather simple structure (e.g. a pure tone). Often, the values for 'pitch' are identical in speech production, transmission, and perception. But the frequency of the quasi-periodically vibrating vocal folds might be attenuated by the vocal tract shape, and a higher harmonic of the F0 can be amplified, which appears as clear periodicity in the acoustic signal and is perceived as pitch. Consequently, pitch and fundamental frequency are not always the same.

As argued in Chapter 2, what is important for speech communication is what is perceived as pitch and not what is produced as fundamental frequency.

Voice source and vocal tract filter information are not linear independent and they are perceived as one signal, rather than being decomposed into separate components. This does not lessen the impact of Fant's (1960) source filter model as an adequate way to describe speech production; but it states that perception is not simply an inversion of the production process. From this it follows directly that the motor theory of speech perception (Liberman & Mattingly, 1985) is untenable, because it predicts the perception of the intended gestures that generated a speech sound. The view that source and filter information are perceived together as one entity is compatible with the invariance theory of speech perception (Blumstein & Stevens, 1981), because here all information in a speech signal is used as an integrated percept, without considering how a sound might have been produced. Finally, the proposed view of pitch perception is also in accordance with the auditory theory of speech perception (Seneff, 1985), because here the acoustic signal is perceived as one percept which is represented in two different ways simultaneously. This thesis argues for an integrated

[1] University of Konstanz, Germany

perception of the speech signal in the auditory periphery for the perception of pitch. Voice source and vocal tract filter information together are used to give the perception of the pitch of a speech signal. The inner ear ultimately converts acoustic energy into electrical energy and warmth, and the temporal pattern of the energy transported by the sound wave (temporal energy distribution, TED) leads to a pitch percept.

Chapter 2 gives an overview of the history of more than 150 years of psycho-acoustic research on pitch perception that is marked by arguments for and against either the rate theory, in which pitch is represented by the firing rate of the neurons of the auditory system, or the place theory, in which pitch is represented by the place of the firing neurons on the basilar membrane in the inner ear. (The rate theory is also called temporal, time, timing, phase-locking, or temporal-rate theory and the place theory is also called place-rate or spatial theory.) The most important arguments against the rate theory are the insensitivity of the ear to phase relations between components (Wightman, 1973b) and the ability to perceive pitch from components presented in both ears separately (Houtsma & Goldstein, 1972). Regarding the first, this thesis argues that the mechanics of the inner ear disturbs phase relations between components of a complex signal at the basilar membrane. Consequently, arguments based on the shape of the acoustic waveform are not compelling enough to explain the pitch perception. With respect to the second argument, this thesis proposes a mechanism similar to that in spatial orientation to explain Houtsma and Goldstein's findings.

To investigate the question that guides listeners in pitch perception - of whether it is the harmonic structure of a signal or its temporal pattern - and related to this, whether it is the rate or the place that defines the perceived pitch, a series of four perception experiments were conducted (Chapter 3). In these experiments speech signals were used in which F0 was different from the periodicity frequency. The first experiment was a forced choice experiment, in which the subjects had to decide whether or not a speech signal stretch of 100 ms had the same pitch as a subsequently presented pure tone. Most often, the subjects decided that the periodicity frequency matched the frequency of the ambiguous speech signals and not F0. The subjects showed a lot of variation in their reactions, suggesting uncertainty in their judgements, which might be an effect of the task of comparing short stretches of real speech signals to perfectly periodic tones. In the second experiment, the speech signals for comparison were made perfectly periodic and the subjects only had to decide whether speech signal and tone were similar in pitch. The subjects chose again the periodicity frequency most often for the critical signals with a slightly improved performance, showing that using artificially periodic speech signals instead of 'real' speech signals did increase the performance of the subjects slightly but does not change their reactions in principle. In the first two experiments the subjects made more consistent decisions when the pure tone was played after the speech signal.

The third experiment was a self-paced matching experiment, where subjects had to adjust the frequency of a pure tones to the frequency of a perfectly periodic speech signal. Here again, the subjects most often selected the periodicity frequency and hardly ever the fundamental frequency. In this experiment, many subjects had problems matching the frequency of the speech signals with pure tones; some subjects could repeatedly match signals and tones with high precision, while others performed very poorly. In the fourth experiment, the matching tone was a 12-tone complex with a 6 dB/octave roll-off. In this experiment, the subjects who could match frequencies with high precision most often chose the periodicity in a speech signal. But the other listeners used the timbre to match the pitch of two signals, and used neither F0 nor periodicity. Even in the fourth experiment, where the experimental set-up was designed to guide the subjects to select F0, the subjects most often selected the

frequency given by the periodicity and not by F0. Additionally, it appeared that subjects in the experiments behave differently, namely that subjects who are able to select a frequency consistently chose other frequencies than less consistent subjects. This raises the question of whether the results in pitch perception experiments can be directly applied to everyday pitch perception.

The conclusion of the first half of the thesis (Chapters 1 to 3) is that pitch in speech perception is best defined on the basis of the temporal structure of a signal, independent of the way the signal has been generated. In particular, it is not the vibrating vocal folds alone that define the perceived pitch of speech signals, but it is the source and the filter together. As a consequence, the often observed 'problem' that pitch meters measure the frequency of a higher harmonic if F0 is attenuated, is actually not a 'failure' of the device or algorithm, but might reflect the actual percept of listeners.

The second half of the thesis (Chapters 4 to 6) deals with the question of how pitch can be measured, where 'pitch' is now understood as the periodic frequency of a signal. In Chapter 4, principles to measure the pitch of a speech signal are critically discussed, in Chapter 5 the TED pitch meter is described, and then evaluated in Chapter 6.

In Chapter 4, articulatory based pitch meters are discussed in terms of electroglottograph (EGG) and accelerometer devices. The principles of speech signal oriented pitch meters were presented with a data reduction method (Schäfer-Vincent, 1982; 1983), which operates in the time domain, and a cepstrum method (Noll, 1967), which operates in the frequency domain. The subharmonic sieve algorithm (Hermes, 1988) represented a speech perception oriented pitch meter that operates in the frequency domain. Autocorrelation and average magnitude difference function were used as examples for hybrid algorithms.

The TED pitch meter presented in Chapter 5 is a time domain algorithm that reduces the speech signal into a sequence of 'needles', which are a coarse representation of the temporal energy distribution in the acoustic signal.

These needles are then tested by 'logical filters' in subsequent steps to yield a quasi-periodic train of needles. Finally, this train is converted into a pitch contour. The conversion of the waveform to the needles in the initial step of the algorithm reduces the influence of phase relations and reduces random background noise - two of the major problems for other time domain algorithms. A further aspect of this transformation is an efficient data-reduction, improving the speed of the algorithm. The algorithm first computes the area between the waveform and the zero line between two reasonable zero crossings, that is, crossings of the zero line that are not incidental spikes. This area is a rough estimate of the energy of the signal between two zero crossings and the temporal distribution of the resulting 'energy' values are examined in the subsequent processing steps. These steps are tests on the amplitude and periodicity pattern of the reduced signal which are performed above and below the zero line independently, inspecting the signal left-to-right and right-to-left. Finally, the two traces above and below the zero line are combined and sequences of periodicity are taken as the pitch of the signal. The algorithm is very fast and delivers a pitch contour for a wide range of speakers and conditions.

Writing a fast pitch determination algorithm is not a problem, the question is: are the determined pitch values correct? To examine this question, a formal assessment of the TED algorithm is presented in Chapter 6. After reviewing the three major reports about pitch meter assessment (Rabiner et al., 1976; McGonegal et al., 1977; Viswanathan and Russell, 1984) a set of error criteria are developed. The error criteria are similar in the three reports, but the definition of a nearly correct frequency match between reference data and pitch meter, and the definitions of the error rates, are

refined in this thesis. In addition, a mixture of different reference signals was constructed to evaluate the TED algorithm. The set of reference signals included recordings made in sound treated rooms, a concert hall, and living rooms. The algorithm was first compared to EGG signals (here, only signals where F0 and periodicity are identical were used), then to human labelers, and finally to other pitch meters. One outcome of this comparison is that the TED algorithm performs pitch estimations at a very high quality level, but it has some tendency to find pitch values in voiceless regions too. The algorithm handles male and female voices equally well, and is fairly insensitive to background noise. By using a rough energy estimation, instead of the waveform amplitude, sensitivity to background noise and phase effects is resolved, thus overcoming two of the major problems for time domain algorithms.

Finally, in Chapter 7, the results of the thesis are discussed and directions for future research and for improvements of the TED algorithm are pointed out. One of the major findings of this research is that it there is no separation of voice source and vocal tract filter in pitch perception in speech, and that the F0 (e.g. as measured with an EGG) need not be the same as the perceived pitch. Subjects perceive both together as one entity. Pitch of a speech signal for an average listener can be best understood as an integral perception of the periodicity in the acoustic signal. Implementing this in a pitch meter operating in the time domain leads to an algorithm that performs fast and very well.