EE 391 Special Report (Autumn 2004)
Pitch Perception: Place Theory, Temporal Theory, and Beyond

Advisor: Professor Julius Smith

Kyogu Lee
Center for Computer Research in Music and Acoustics (CCRMA)
Music Department, Stanford University
kglee@ccrma.stanford.edu

Abstract

Two competing theories on pitch perception are reviewed with brief history and several significant works that contributed to building such theories. Fundamental concepts behind these two theories - place theory and temporal theory - are briefly described, and the further steps their successors took are presented, followed by possible future directions.

1 Introduction

Pitch is one of the most important attributes of audio signals such as speech and music. In speech, pitch can greatly improve speech intelligibility and thus can be very useful in speech recognition systems. Sound source separation is another application where pitch information is critical especially when there are concurrent sound sources. In musical application such as automatic music transcription, pitch is indispensable since it directly corresponds to the height of the musical notes. There are many other applications where pitch information is of great use such as pitch-shift audio effects or parametric audio coding, to name a few.

Pitch perception is a very complex sensory phenomenon which involves a lot of sciences such as physics, psychology, psychophysics, psychoacoustics, physiology, and neurological science, and therefore there is no unitary theory or model capable of explaining all the processes how human perceives a pitch. Nevertheless, there have been two long-lasting theories in rivalry on pitch perception, either of which most experts agreed to with their own variations. One is a place theory, and the other is a temporal theory. Although they have been mainstreams of pitch theories, none of which can explain all the phenomena represented by various hearing experimental data. Some phenomena that a place theory fails to explain are easily explained by a temporal theory, and vice versa. Hence it would be fairer to call them complementary than competing, which may be proved by later hybrid models.

2 Place Theory

The place theory has a long history which may hark back to the days of Helmholtz (von Helmholtz 1954) although most ideas can be traced back far beyond him.(de Cheveigné 2004). According to his resonance-place theory of hearing, the inner ear acts like a frequency analyzer, and the stimulus reaching our ear is decomposed into many sinusoidal components, each of which excites different places along the basilar membrane, where hair cells with distinct characteristic frequencies are linked with neurones. He also suggested that the pitch of a stimulus is related to the pattern of the excitation produced by the stimulus along the basilar membrane. For a pure tone, the pitch generally corresponds to the position of maximum excitation; for a complex tone with many spectral peaks, it is more complicated. Among the supporters of the place theory, Goldstein, Terhardt, and Wightman are best known with all the differences in detail.

The biggest problem of the place theory is that it fails to identify the pitch of a stimulus with missing fundamental. According to Helmholtz's theory, it is impossible to perceive a pitch when there is no spectral peak at the position along the basilar membrane which corresponds to the frequency of the pitch. Nevertheless, there is a pronounced pitch at the missing fundamental. Schouten (Schouten 1938) did extensive experiments on this problem using a periodic pulse where he discovered a low pitch associated with high harmonic components of a stimulus. He called this percept a “residue pitch”.

While Helmholtz's theory was not able to explain how we perceive a residue pitch, he did suggest a few options without renouncing his theory to solve a missing fundamental problem - the first was the *nonlinear distortion* invoked in the inner ear; the second was the concept of *unconscious inference*, which later became a foundation of a *pattern matching* model.

Wightman (Wightman 1973) formalized a mathematical model, the so-called "pattern-transformation model" of pitch based on auditory pattern recognition. In his model, Wightman used the term "pattern" to refer to two dimensional distribution of neural activity - *place* and *amount*. Different places in the pattern represent individual or groups of nerves, and the amount given by the pattern indicates the activity of these nerves. Wightman hypothesized the power spectrum of the waveform might represent "peripheral activity pattern", and used it as the initial pattern in his model. This pattern then undergoes Fourier transformation, which yields another pattern similar to autocorrelation function of the stimulus. Finally, pitch information is extracted by finding the places in the transformed pattern where activity is maximal. This model solved the problem of phase insensitivity which was not the case of such temporal models as "peak-picker" or "fine-structure".

Goldstein (Goldstein 1973) introduced a central processor theory, according to which the central processor is a recognizer of spectral patterns supplied by frequency analyzers. Recognition is accomplished by finding the best matching stored pattern or template. In more detail, the central processor makes an optimum estimate of the unknown fundamental of the stimulus using maximum likelihood estimation. The important assumption made here is that the stimulus frequencies are unknown successive harmonics of the unknown fundamental. To illustrate how his theory works, consider a complex tone which has components at 1840, 2040, and 2240 Hz. For such stimulus, the central processor would find a good match of a harmonic complex tone with a fundamental of 204 Hz since its 9th, 10th, and 11th harmonics are at 1836, 2040, and 2244 Hz. In fact, as shown in Schouten *et al.* (Schouten, Ritsma, and Cardozo 1962), the perceived pitch is close to that of a 204 Hz pure tone. However, in such non-harmonic stimulus, there is a pitch ambiguity, and the weaker pitch percepts are also found around 185 Hz and 227 Hz. Goldstein's theory is also able to identify these weak pitches by finding two other matches; one to be the 8th, 9th, and 10th harmonics of a 226.7 Hz fundamental, and the other to be the 10th, 11th, and 12th harmonics of a 185.5 Hz fundamental.

Terhardt (Terhardt 1974) suggested that the essential principle in explaining the phenomena of pitch perception is the distinction between *spectral pitch* and *virtual pitch*. According to him, for example, the pitch of a pure tone is a spectral pitch while that of a complex tone is virtual pitch. In addition, Terhardt presented two modes of pitch perception; *analytic mode* results in spectral pitch, and *synthetic mode* results in virtual pitch. Although there are two distinct kinds of pitch, both are derived from spectral cues. In his model, the spectral cues are first extracted from the stimulus, and the virtual pitch is assumed to be a subharmonic of a dominant partial, by which he meant a partial that is resolvable, i.e., which can be heard out from the complex tone. He also suggested to include a learning phase of harmonically rich sounds such as speech since we are innately exposed to such sounds from birth.

All of the models described above depend on the spectral resolution of individual components in the stimulus. As Terhardt mentioned, the residue pitch or virtual pitch will be perceived only when some of frequency components are resolved or "heard out" from the tone complexes. Therefore, the place theory fails to identify the pitch of complex tones whose harmonics are too close to be resolved or there is no well-defined spectral structure in the stimulus such as interrupted noise. This is where the temporal theory wins over the place theory.

3 Temporal Theory

While place theory tries to explain pitch sensation by finding places in the basilar membrane where the excitation by the stimulus is maximal, temporal theory is a time-domain mechanism which is event-based; i.e., it tries to detect the time interval between *events*, which may be peaks or overall envelope of the input waveforms. These events determine the periodicity of the waveform, and the reciprocal of the periodicity is the same as the fundamental frequency. In his residue pitch theory, Schouten(Schouten 1938; Schouten, Ritsma, and Cardozo 1962) mentioned the important role of unresolved high harmonics. According to him, pitch sensation of a complex tone occurs when an interaction of those several unresolved harmonics results in a periodic time pattern of the waveform, and the residue pitch is determined by the periodicity.

Temporal theory like Schouten's can explain phenomena which place theory fails to interpret. The first one is the problem of missing fundamental. Because the residue pitch does not correspond to any of physical sinusoidal components of the stimulus, but is determined by overall time pattern caused by an interaction of several harmonics, the fundamental can be physically absent to invoke pitch sensation. In addition, pitch sensation of interrupted noise with no spectral peaks is also explained by the temporal theory. However, the role of unresolved harmonics, which was essential in the residue pitch theory, turned out to be wrong by the findings of Plomp(Plomp 1967) and Ritsma(Ritsma 1967). In their studies, they found a so-called dominant region covered by the frequency components of the third, fourth, and fifth harmonics, and they proposed that the pitch of a complex tone is determined by the dominant spectral region, where the harmonics are obviously resolvable.

Licklider introduced a method of autocorrelation analysis in his duplex theory of pitch perception(Licklider 1951), the essence of which is that our auditory system employs both frequency analysis and autocorrelation analysis for sensation of pitch. Frequency analysis is performed by the cochlea via an array of bandpass filters, and autocorrelation analysis is performed on the the activity of auditory nerve fibers, resulting in a two-dimensional pattern: characteristic frequency and time lag. The pitch is then extracted from nerve firing patterns by finding a time lag with maximal peaks in the autocorrelation function. Meddis and Hewitt(Meddis and Hewitt 1991a) took a further step to propose a summary autocorrelation function (SACF), which is basically an integration of autocorrelation functions across auditory channels. The highest point of the SACF is used to indicate the perceived pitch, and Meddis and Hewitt argued that many phenomena about pitch perception could be explained with their model including the missing fundamental, ambiguous pitch, the pitch of interrupted noise, inharmonic components, and the dominant region of pitch. A computationally efficient model was later developed by Tolonen and Karjalainen(Tolonen and Karjalainen 2000), where they introduced an enhanced summary autocorrelation function (ESACF), which divides the signal into two channels, below and above 1000 Hz, while there are 128 channels whose center frequencies lie between 80 Hz and 8 kHz in Meddis and Hewitt's model. In spite of this drastic simplification, Tolonen and Karjalainen could demonstrate the model performance to be comparable to other time-domain models using a multichannel analysis.

4 Other Models

Pitch sensation involves so many complex processes that no theory alone can account for all of the experimental data. The fact that a few low resolvable harmonics dominate the pitch percept seems to support a pattern matching theory. It is easy to explain phase insensitivity of our hearing mechanism if we adopt a place theory, although Meddis and Hewitt(Meddis and Hewitt 1991b) suggested a solution to the problem. On the other hand, it is possible to perceive a residue pitch when harmonics are too high to be resolved, and this is in favor of a temporal theory.

Such complementary properties between a place theory and a temporal theory gave birth to a number of related models including hybrid models. Considering that autocorrelation function and power spectrum is a Fourier transform pair, the autocorrelation model formulated by Licklider might be said to be a hybrid model. Srulovicz and Goldstein(Srulovicz and Goldstein 1983) also introduced a hybrid model called a central spectrum model, where the auditory nerve to the brain is given in the form of a tonotopic display of stimulus spectrum, called central spectrum, and the pitch percept is derived from this central spectrum by an optimum pattern recognizer.

A schematic model proposed by Moore(Moore 1977) is similar in concept to a model by Meddis and Hewitt(Meddis and Hewitt 199 where acoustic input is filtered by a bank of bandpass filters, and the filter outputs are transduced to neural impulses. The interspike intervals are then analyzed for each characteristic frequency, combined across the channels, and most prominent intervals are fed to a pitch decision mechanism which selects one interval.

A time-domain cancellation model by de Cheveigné(de Cheveigné 1993 is similar to an autocorrelation model, but is based on a difference function instead of an autocorrelation function, which can be useful for separation of concurrent sound sources with multiple pitches.

Klapuri(Klapuri 2003) took an iterative approach to estimate multiple fundamental frequencies of concurrent musical sounds. In his method, the fundamental of the most prominent sound is estimated using the frequency relationships of simultaneous spectral components, without assuming perfect harmonicity. The spectral envelope of the detected sound is then estimated by applying the spectral smoothness principle. The estimated envelope is subtracted from the mixture, and the procedure is repeated for the residual signal.

5 Discussion

The inability to explain all of the experimental data related to pitch perception with a unitary theory or model led to a view that there might exist two separate pitch perception mechanisms: place or spectral mechanism for low, resolvable harmonics, and temporal mechanism for high, unresolvable harmonics(Carlyon and Shakleton 1994). However, Meddis and O'Mard(Meddis and O'Mard 1997) argued that both types of mechanisms could be embraced in a unitary model using a summary autocorrelation function method. On the other hand, in very recent experiments designed by Oxenham *et al.*(Oxenhan, Bernstein, and Penagos 2004), they succeeded to prove that tonotopic representation is crucial to pitch perception of complex tones by demonstrating the subjects' inability to extract the fundamental frequency from multiple low-frequency harmonics presented to high-frequency regions of the cochlea. This result is strongly in favor of the place theory.

Despite the efforts to explain pitch perception with a single theory or model, pattern matching model and autocorrelation model seem to be two major options for explaining pitch today, and this is possibly the reason why there are a number of variants of these models around. The author is particularly interested in de Cheveigné's view of using the *string* to model our auditory system(de Cheveigné 2004), where he mentioned there is a close relation between the string and autocorrelation. According to him, autocorrelation consists of two basic processes: *delay* and multiplication. A string is, in essence, a *delay* line that feeds back onto itself. Both autocorrelation and string have peaks at the multiples of the period of the input signal, but it is much sharper in the latter case. Therefore, it might be possible to improve the accuracy or the efficiency of the autocorrelation model using the string to model our auditory system.

References

Carlyon, R. P. and T. M. Shakleton (1994). Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *Journal of the Acoustical Society of America 95*(6), 3541–3554.

de Cheveign´e, A. (1993). Separation of concurrent harmonics sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America 93*(6), 3271–3290.

de Cheveign´e, A. (2004). *Pitch perception models in Pitch*. Springer-Verlag. Edited by C. Plack and A. Oxenham, 2004.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America 54*(6), 1496–1516.

Klapuri, A. P. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing 11*(6), 804–816.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia 7*(4), 128–134.

Meddis, R. and M. J. Hewitt (1991a). virtual pitch and phase sensitivity of a computer model of the auditory periphery. i:pitch identification. *Journal of the Acoustical Society of America 89*(6), 2866–2882.

Meddis, R. and M. J. Hewitt (1991b). virtual pitch and phase sensitivity of a computer model of the auditory periphery. ii: Phase sensitivity. *Journal of the Acoustical Society of America 89*(6), 2883–2894.

Meddis, R. and L. O'Mard (1997). A unitary model of pitch perception. *Journal of the Acoustical Society of America 102*(3), 1811–1820.

Moore, B. C. J. (1977). *An Introduction to the Psychology of Hearing*. Academic Press. First edition.

Oxenhan, A. J., J. G. W. Bernstein, and H. Penagos (2004). Correct tonotopic representation is necessary for complex pitch perception. *Proceedings of the National Academy of Sciences 101*(5), 1421–1425.

Plomp, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America 41*(6), 1526–1533.

Ritsma, R. J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America 42*(1), 191–198.

Schouten, J. F. (1938). *The perception of subjective tones in Psychological Acoustics*. Edited by E.D. Schubert, 1979.

Schouten, J. F., R. J. Ritsma, and B. L. Cardozo (1962). Pitch of the residue. *Journal of the Acoustical Society of America 34*(8), 1418–1424.

Srulovicz, P. and J. L. Goldstein (1983). A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. *Journal of the Acoustical Society of America 73*(4), 1266–1276.

Terhardt, E. (1974). Pitch, consonance, and harmony. *Journal of the Acoustical Society of America 55*(5), 1061–1069.

Tolonen, T. and M. Karjalainen (2000). A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing 8*(6), 708–716.

von Helmholtz, H. L. F. (1954). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York: Dover. English translation of 1863 (German) edition by A. J. Ellis.

Wightman, F. L. (1973). The pattern-transformation model of pitch. *Journal of the Acoustical Society of America 54*(2), 407–416.