Proc. of the National Conference on Communications, I.I.T. Kanpur, January 2001.

Applying Perceptual Distance to the Discrimination of Sounds

Preeti Rao

Advanced Centre for Research in Electronics,
Indian Institute of Technology, Bombay

Abstract

In sound compression and synthesis, it is valuable to have objective distance measures which can predict the perceived dissimilarity of two sounds. Furthermore it is desirable also to be able to estimate the extent of the dissimilarity or the quality difference between sounds. Since it is the perceived difference that needs to be quantified it can be expected that measures that are derived from models of hearing will provide the best predictions. An overview of the topic of auditory modelling is presented. An objective measure based on a distance metric applied to auditory excitation patterns is described. This perceptual measure is compared with a traditional spectral distortion measure in the prediction of the discrimination of spectral envelope distortions of vowel-like sounds. The results of a subjective experiment indicate that the perceptual measure is significantly superior.

Introduction

In the fields of audio signal compression, audio content retrieval, and synthesis of speech and musical sounds, the problem arises of predicting whether two sounds are perceived as different and of how to express the extent of the perceived difference in quality. Objective distance measures for detecting an audible difference in quality, and quantifying it, are important in these areas. Such measures can serve as convenient substitutes for subjective quality measurement, and can also be applied in analysis-by-synthesis algorithms where the difference between a reference sound and the corresponding synthesized sound needs to be estimated. Traditionally used objective distance measures, such as signal-to-noise ratios and spectral distances, are derived directly from differences in the waveforms or in the power spectra of the reference and test signals. However, because it is the perception of the distortion that needs to be quantified, it is natural to expect that measures derived from models of the auditory system will provide the most accurate predictions.

While a sound can be described in terms of the physical properties of the acoustic signal such as its intensity and frequency spectrum, the corresponding sensation perceived by the listener is described in terms of psychoacoustical attributes such as loudness, pitch and timbre. A computational model of auditory perception would ideally predict these psychophysical attributes from the acoustic waveform or spectrum. To detect differences in sound quality then, a reasonable approach would be to compute the appropriate "internal auditory representation" of the reference sound and of the test sound, and compare the two.

In the remaining part of the paper we give an overview of auditory modelling and describe a perceptual distance measure for estimating the quality difference between steady vowel-like sounds which differ in spectral envelope. This measure is evaluated in a subjective experiment and compared with the traditional log spectral distortion metric.

Models of perception

Auditory models are typically based on a combination of data from physiological studies and psychoacoustical experiments (which attempt to correlate physical properties of sound with the sensations as reported by subjects). The auditory system includes the auditory periphery (ears) and central processing (nervous system). Acoustic vibrations of the eardrum are transmitted by the middle ear to the basilar membrane of the cochlea. If the sound is a pure tone of given frequency, the resulting oscillations are localised to a fixed spatial region of the basilar membrane and activate the hair cells and neurons associated with this region. The spatial position of this "resonance" region depends on the frequency of the tone. As the frequency is increased, the centre of the corresponding resonance region moves from the apex of the basilar membrane toward its entrance with the relation between frequency in Hz and spatial location in millimetres being roughly logarithmic. Thus the information about frequency is "coded" in terms of the spatial location of the activated neurons. The spatial *extent* of the resonance region is independent of the frequency of the tone and corresponds to a fixed length along the basilar membrane. The neural activity at any spatial location is therefore influenced by all frequency components that lie within a certain bandwidth which corresponds to a fixed spatial width about that location. This frequency selective behaviour of the inner ear is typically modeled as a bank of overlapping, bandpass filters with centre frequencies equally spaced along the length of the basilar membrane. The bandwidth of the auditory filter, known as the critical bandwidth, corresponds to a fixed spatial width of the basilar membrane and hence increases roughly logarithmically with centre frequency.

While most of the data related to auditory models comes from physiological measurements performed on animals, psychoacoustical experiments have provided much data on the psychophysical estimates of pitch and loudness (Roederer, 1995). Most of these studies support the "place theory of hearing", that is the spatial pattern of neural activity generated by basilar membrane mechanics encodes all the information necessary to explain the subjective sensations related to pitch, loudness and timbre. (There are, however, certain phase-sensitive effects such as the perception of beats that suggest an analysis of the temporal pattern of vibration.) The specific distribution in space and time of auditory neural activity is a suitable definition for the "auditory internal representation". Auditory excitation patterns, the term used to describe the distribution of basilar membrane excitation level per critical band, were first proposed by Zwicker and Scharf (1965) as suitable auditory internal representations. These are computed from the power spectrum as the output of the auditory filters with centres distributed uniformly on a critical band scale. While this internal representation is generated by the peripheral auditory system, the extraction of auditory features and other higher order processing is attributed to the (not yet fully understood) central processing of the brain. It can be assumed, however, that perceptual *discrimination* (under optimal listening conditions and using well-trained subjects) depends largely on the resolution properties of the auditory periphery, and should be predictable by any good model of peripheral auditory processing.

Perceptual distance measures for vowel quality

The steady portions of vowel sounds belong to the general class of sounds which can be represented as a set of harmonic components at multiples of a specified fundamental frequency. We consider those distortions of the sound that can be described by changes of the spectral magnitudes of the harmonic components. Sources of this type of distortion are, for example, filtering by a non-uniform gain transfer function and the inaccurate modelling of the spectral envelope, for instance, in linear predictive synthesis. It is of interest to predict whether the modifications give rise to discriminable changes in perceived quality, and if so, to quantify the extent of perceptual degradation.

Typically the log spectral distance (LSD), given by the root of the mean squared difference calculated between the log spectral magnitudes of the reference and test signals, is applied to quantify the spectral distortion (Quackenbush et al,1988). The more direct incorporation of auditory processing knowledge is seen in the work of Plomp (1976) who addressed the problem of prediction of perceptual differences for steady vowel sounds in the following manner. The spectral levels of the input sound power spectrum were summed over 1/3-octave bands, approximating the critical bands of the auditory system, to obtain a spectral representation more closely matched to auditory excitation patterns. The quadratic distance between the reference and test signal representations was used to predict subjective quality differences in a set of steady sounds. Auditory excitation patterns, as computed from a model of peripheral auditory processing, were used by Gagné and Zurek (1988), who investigated resonance-frequency discrimination of single vowel formants. The difference in the excitation patterns of the reference and modified signals was used to derive a distance measure given by either the single, largest magnitude difference (single-band model) or by the appropriately combined differences across bands (multiband model). In the present work, we consider a perceptual measure, also based on excitation patterns, but computed from a recently proposed and relatively detailed model of peripheral auditory processing (Moore and Glasberg, 1987).

The excitation pattern of a sound is calculated as the output of the auditory filters of the inner ear which model the frequency selectivity of hearing at specific centre frequencies. The input signal power spectrum, specified by the frequencies and power spectral levels in dB SPL of its components, is first attenuated by the transfer functions for the outer and middle ear to obtain the effective spectrum reaching the cochlea. Auditory filter shapes are characterised as being rounded exponentials with parameters that control the filter selectivity. The frequency selectivity depends both on the centre frequency of the auditory filter and the input stimulus level. With increasing input level the lower slope of the filter becomes shallower. The contribution of each stimulus component to the excitation pattern is calculated with a filter shape particular to that component. Thus to calculate the excitation level corresponding to the output of a given auditory filter, the input power spectral components are each weighted depending on their level and distance from the filter center frequency and combined additively. This is repeated for all filter center frequencies, spaced at uniform intervals on a critical band scale, in the range of 50 Hz to 15 kHz. We thus obtain the complete excitation pattern as a density, i.e. in dB SPL per critical band.

We define our perceptual distance measure (henceforth referred to as PDM) as the quadratic distance (in dB SPL) between the reference and test signal excitation patterns computed as discussed above.

Experimental validation

A subjective experiment was carried out to validate the performance of the perceptual distance measure in the prediction of audibility discrimination thresholds for arbitrary modifications of the spectral envelope of steady harmonic complexes. Prediction of the discrimination threshold is a part of the larger problem of quantifying the perceptual effect of a distortion of the signal. A representative set of modification conditions were chosen for the experiment, distributed over the spectra of two simulated steady vowels, /a/ and /i/. A good measure is expected to produce the same threshold values for distinct conditions, at least for an individual subject.

The reference sound spectra were derived from the amplitude spectra of the vowels synthesised by the cascade combination of an LF model glottal source and a formant filter based on the linear prediction coefficients. A constant overall level of about 55 dB SPL was maintained. The test set of modifications was chosen in a way to encompass distinct types of gain changes of the spectral envelope. The amplitudes of the harmonics were modified by multiplication with a factor close to one. When more than one harmonic was modified, each harmonic was multiplied by the same factor. Both, localised as well as relatively spread modifications (e.g varying the overall spectral tilt) were tested. Figure 1 depicts the set of stimuli and modifications by indicating which harmonic components are affected in each of the conditions. A fundamental frequency of 220 Hz was used. In each of the conditions, the spectral amplitudes were modified in small steps. Reference and modified sounds of duration 300 ms were generated as the sum of harmonics with the specified amplitudes and random phases.

Four normal-hearing subjects participated in the experiments. The stimuli were presented binaurally over headphones at a level of approximately 55 dB SPL to subjects seated in a sound-proof booth. For each condition we determined the amount of spectral magnitude change at which the subject was just able to discriminate between the reference sound and the modified sound. This was achieved by the use of a 3-interval forced-choice adaptive procedure in which the subject is required to identify the (single) interval which contains the modified sound. The amount of spectral envelope distortion at which the subject achieves 70% correct responses is considered to be the threshold condition of discriminability for the corresponding distortion type.

For each condition and subject, the objective distance (both LSD and PDM) between the reference sound and the modified sound at threshold were computed. The threshold distance values, averaged across the four subjects, are plotted versus condition number in Figure 2.

Discussion

A desirable property for any objective measure is that the value obtained at the threshold of discriminability should be approximately constant for the different spectral modifications. We see from Figure 2 that both the LSD and the PDM for the distinct types of spectral envelope distortion are spread in a narrow range about their mean values. To compare the performance of the two measures we need to quantify the dispersion of each as observed in the experimental data. A typically used estimate of dispersion is the "relative variation" defined as the standard deviation divided by the mean. It describes the extent of dispersion and is invariant to scale changes in the definition of the objective measure. From the data of Figure 2 we obtain the following values for the relative variations of threshold levels across the 10 conditions in the experiment:

LSD = 0.49 ; PDM = 0.21

Based on the relative variation, the perceptual measure is a significantly more accurate predictor of the subjective data. An intuitive feeling for this result can be obtained by studying the difference in the underlying assumptions in the definitions of the two objective measures. The log spectral distortion is premised on the fact that the perceptual resolvability of intensity is proportional to the ratio of the two intensities. Equivalently, a magnitude change of a smaller magnitude spectral component will be more easily detected than that of a larger component. However, the LSD does not take into consideration the "masking" effects of the neighbouring frequency components. A large component in the close proximity can greatly reduce the detectability of distortion in a spectral valley. An example of this behaviour in evident in Figure 2 in which the LSD shows a relatively large increase in value going from Condition 1 (spectral peak) to Condition 3 (spectral valley with masking effects). The PDM, on the other hand, is based on a model of hearing and implicitly accounts for masking.

Considering more specifically the problem of predicting perceived distance in vowel quality, there are other types of modifications of the spectral envelope which are relevant. These include changes in formant frequency and bandwidth, and phase relations among harmonics. These have not been considered here.

It would also be of interest to explore whether the PDM can predict quality differences which are clearly audible, that is they are above the threshold of detection. While it is likely that quality differences near the threshold of audibility can be quantified by the excitation pattern based PDM, it is not clear whether the larger supra-threshold quality differences will be estimated correctly. Subjective judgement of quality difference (e.g. determining whether one pair of sounds is more similar than another pair) has been observed to depend on the comparison of various higher order features rather than on a simple comparison of "raw" auditory excitation patterns.

Acknowledgements

Parts of this work were carried out during the author's visit to IPO, Eindhoven, in close collaboration with R. van Dinther, R. Veldhuis and A. Kohlrausch.

References

Gagné, J. P. and Zurek, P.M. (1988). Resonance-frequency discrimination. *Journal of the Acoustical Society of America, 83,* 2293-2299.

Moore, B.C.J. and Glasberg, B.R. (1987). Formulae describing frequency selectivity as a function of frequency and level and their use in calculating excitation patterns. *Hearing Research, 28,* 209-225.

Plomp, R. (1976). Aspects of tone sensation. Academic Press Inc. London (LTD).

Quackenbush, S. R., Barnwell, T.P. and Clements, M.A. (1988). Objective Measures of Speech Quality. Prentice Hall, Englewood Cliffs, New Jersey.

Roederer, J.G., (1995). The Physics and Psychophysics of Music. Springer-Verlag, New York.

Figure 1: The spectral distortion conditions used in the subjective experiment.
The encircled points indicate which harmonics were scaled up or down.
Not shown are the two modifications of changing spectral tilt (conditions 5,6)

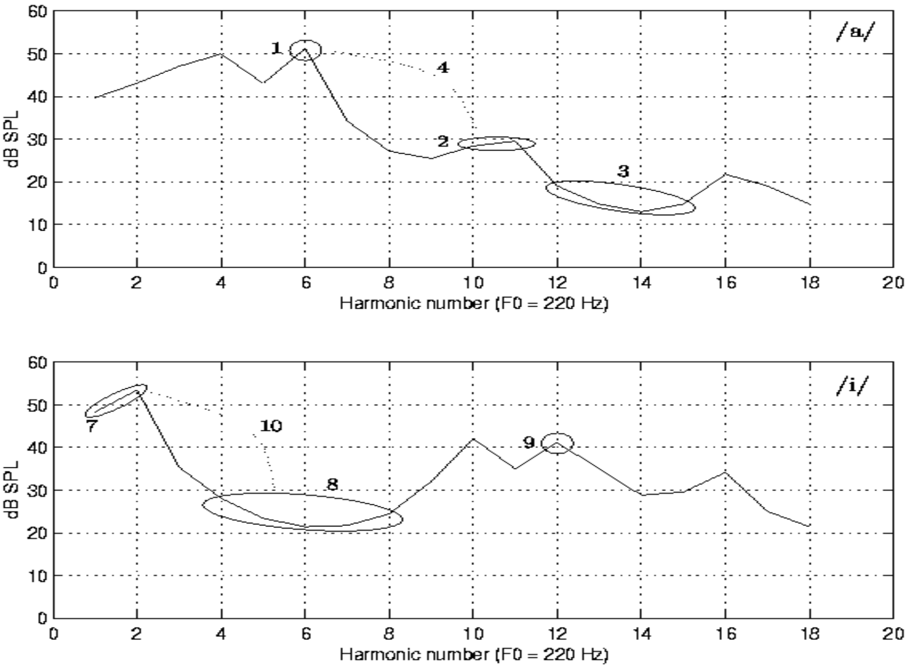
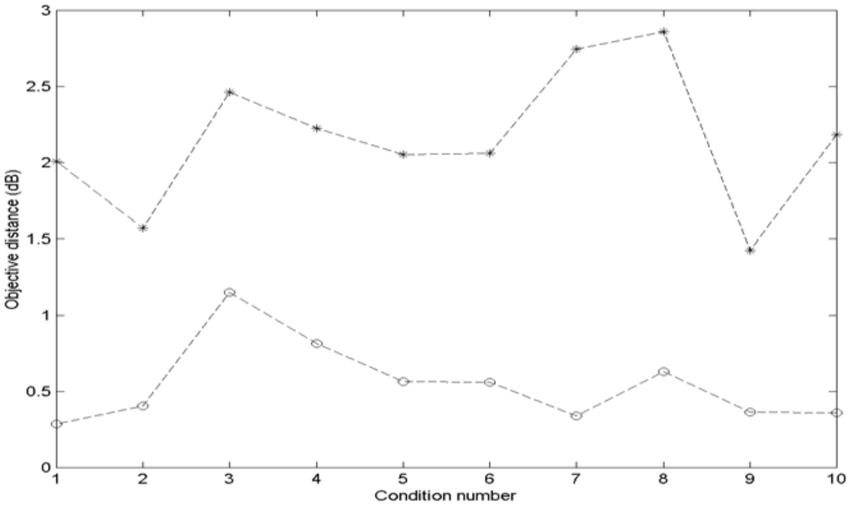


Figure 2: Objective distance calculated at threshold for each condition.
' * ': PDM; ' o' : LSD


Objective distance (dB)
Condition number