Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney, 2002
C. Stevens, D. Burnham, G. McPherson, E. Schubert, J. Renwick (Eds.). Adelaide: Causal Productions.

SPEECH AND MUSIC, ACOUSTICS AND CODING, AND WHAT MUSIC MIGHT BE 'FOR'

Joe Wolfe

School of Physics, The University of New South Wales, Sydney
J.Wolfe@unsw.edu.au

ABSTRACT

In both music and speech, the perception of different subsets of acoustical features is categorical, and the categorically perceived features are most extensively notated. However, the way in which these features are used to encode different elements of the signals in music and speech is quite different, and in some ways complementary. These simple and general observations may offer some insight into the larger and speculative questions about the nature of music. One such speculation, to be entertained at the conference, is that music is an auditory game.

The presentation of this paper will be a structured discussion and debate involving several researchers and musicians. The present paper serves as an introduction.

1. INTRODUCTION

Acoustically, music and speech are fundamentally similar. Both use sound, and so are received and analysed by the same organs. Many of their acoustical features are similar, although used in different ways. One purpose of this paper is to compare and to contrast them.

Functionally, speech and music are fundamentally different. This is partly because they encode information of a different sort. They also encode it in fundamentally different ways. These differences are related to some interesting but difficult questions about music, and a second purpose of this paper is to use the discussion of acoustic features and coding to see what they might tell us about these questions.

Oversimplifying considerably for the sake of the argument, one could say that speech usually has an explicit and often denotative meaning, upon which many listeners can agree. Music usually does not [1]. This difference renders speech more obviously useful. Why do humans have speech? The evolutionary biologist can point to the potential survival and mating advantages of speech. Why do humans have music? The professional musician might argue that it confers advantage in mating, and to a lesser extent, in survival. But was it always so? Once humans have an appreciation for music, we can find uses for it. But why do we appreciate and have the capacity to perceive musical elements in the first place?

Attempts to address such questions are inevitably speculative. However, one might expect to gain some insight relevant to such speculations by comparing and contrasting the acoustics and the coding in music and speech. Other disciplines, of course, also have much to offer in addressing such broad questions. For that reason, the conference presentation of this topic will be a structured discussion and debate, and will involve a number of researchers and musicians with different backgrounds and views. The present paper aims to provide some relevant background and to pose some questions.

An important class of music—singing—is produced by the same apparatus that produces speech. Some important classes of musical instruments—bowed strings, woodwinds, lip reeds and organs—have a strongly non-linear interaction involving between a control oscillator and a resonator and so share with the voice the capacity to produce inherently harmonic spectra [2]. The comparison in this paper is restricted to such instruments. The music discussed will be restricted to simple monophonic melody, which, apart from being easy to discuss in a short treatment, is probably one of the earliest forms of music.

2. ACOUSTICAL COMPARISON

In the first stage of the comparison, we consider only the acoustic features of the signal and not the coding: features that we can observe in the acoustic pressure as a function of time, its spectrum, or some combination of the two.

For most of its duration, normal speech in most languages consists of quasi-periodic signals. The vowels (a,e etc), approximants (l,r etc), nasals (m,n) and some of the signal required to identify plosives, (b,g etc) are voiced: they involve quasi-periodic vibration of the vocal folds. The spectra of quasi-periodic signals are of course quasi-harmonic: the spectrum is dominated by a relatively small number of narrow bands. The spectral envelope usually has features on a larger scale, whose peaks are called formants (which are produced by resonances in the vocal tract).

For the majority of its duration, the signal of the simple, monophonic melody consists of quasi-periodic signals: the sustained part of the musical notes. They are produced by the quasi-periodic vibration of an excitatory mechanism (the bow-string interaction, reed, air-jet, lip-reed or vocal folds). Consequently, the spectra of these parts of the signal are quasi-harmonic. The spectral envelope usually has large-scale features such as formants (produced by eg. the resonances of the bridge or the air in the body of a violin, or the cut-off frequencies of bells or tone-hole lattices in woodwinds).

The speech signal is frequently interrupted by short silences. These are produced by plosives, and it is interesting to note that they occur, more often than not, within words or syllables rather than between them. (In the phrase 'words and syllables', normally pronounced, there are just three silences,

ISBN 1 876346 39 6

which occur during the 'd's and the 'b'.) The starting and ending transients associated with these silences involve signals that are less periodic. They usually have some broad band 'noise' and involve the gradual development or disappearance of the harmonic structure. [3,4].

Melodic music is also usually interrupted by silences of various lengths, and starting and finishing transients. The transients usually have some broad band signal, some non-harmonic components and rapidly varying spectra.

In both cases, the spectrum and its envelope vary in time. The spectra are usually most harmonic during sustained signals. In general, loud signals have proportionally more energy in high frequencies. Both signals usually have some jitter or other fine time structure in the harmonic sections, without which they sound artificial or synthetic.

The similarities are in some cases due to fundamental physics such as non-linear excitation, in some due to the fact that I have omitted plucked strings and percussion instruments, and possibly also that musical instruments have 'evolved' (been selected) to share features with the voice.

Consonants have bursts of broad band 'noise' either alone (unvoiced, eg s,sh,p,t) or superposed on a periodic signal (voiced: z,j,b,d). Broad band components of the starting transients are also observed in the spectra of wind and string instruments, and are important clues in the identification of the instrument (ie. of identifying timbre). The fact that they are (nearly) always present on the start of an articulated note has the effect that, on some instruments, they are often not noticed by listeners, and even experienced players, until attention is drawn to them.

These shared acoustic features are largely analysed in the cochlear itself or in low level processing centres and so, at this level, the hard- and soft-ware used for processing are very largely shared by speech and music.

One difference is the short-term stability of the frequency components in the harmonic sections in music. In most speech, the equivalent of pitch varies continuously, whereas the pitch of individual notes in music is relatively stable[1].

Another difference is the stability over time of formants. In speech, these vary from one phoneme to the next. Further, during one phoneme, their frequencies are usually varying somewhat smoothly towards the values they will take in the next phonemes. In music, on the other hand, formants such as those due to the resonance of a violin bridge, or to other acoustic properties of the instrument, may not vary much at all. (Exceptions are brass with 'wah' mutes—whence the name—and electric guitars with effects pedals.)

[1] Some of this difference between the two is perceptual, however, rather than acoustic. Portamento and other pitch changes may be relatively important in music, and the apparent stability the result of categorical perception. In speech, individual phonemes may have relatively stable frequency components, particularly if the speech is slow.

To discuss other important differences between speech and simple music, it is more helpful to consider the coding and larger structures in the signals.

3. CONTRASTING CODING

Coding refers to the way in which information is transmitted in a signal. Discussion of the information present in a signal is complicated by the fact that information of different sorts has different value, sometimes comes from different sources, and may be interpreted in different ways. In the case of music, some information is provided by the composer, some by the performer (the same person, if improvised), some by the instrument and some by the various effects of the acoustic environment of the performance. If an actor performs a written text, similar divisions may be made, where in this case the vocal apparatus is the instrument.

Let us call the information that would be notated by composer/writer the textual information. The importance that many people attach to this information belies its small size. Thus we refer to a few pages of notes as 'being' a Bach 'cello suite[2], or a slim volume of text as 'being' a play by Shakespeare. The player/actor adds what we may call performance information. (This information, among other things, allows us to distinguish two different players and their musical training.) The instrument, under the control of the player, adds what we shall call carrier information: it provides a complex waveform whose properties are controlled and modulated by the player, in a way that is more complex than, but which may be compared with, the way in which a radio broadcast signal modulates the carrier wave. The output signal is convoluted with functions of the performance space to give the signal received by the listener.

Both speech and music are transcribed with quantised or digital codes[3], which very largely correspond to features of

[2] Because we notate primarily the categorically perceived, quantised variables, written music contains relatively little information. Large works can be stored in digital files, which are very much smaller than those of a digital recording of a performance. The data compression of inherently analogue variables such as sound pressure is both limited and complicated. One could however say that the words 'Cello' and 'Bach' at the top of a piece of music are an extremely efficient compression of a great deal of information: a suitably sophisticated receiver and coding scheme (including a cello, the skill to play it and knowledge of performance technique and traditions) can construct an information rich analogue signal, even if the reconstruction is, at the level of the sound wave (the 'carrier' above) and in some levels of performance, quite different from performance to performance.

[3] Whereas digital electronics uses primarily binary coding, music and language use many-valued quantisation. Languages usually have several dozen phonemes, and music typically several dozen pitches and note lengths. However, whereas digital electronics usually quantises just the voltage, phonemes and notes involve more than one perceptual dimension. Further, there is an inevitable compromise between transmission speed and the size of the quantisation.

ISBN 1 876346 39 6

the acoustic signal that are categorised in perception [5,6]. In performed speech and music, other features are present that do not appear to be subject to categorical perception (e.g. timbre components in music, pitch in Western speech). The use of the acoustical features described above are very different in music and speech.

The advantages of a digital over an analog signal are well known in other contexts: reduced susceptibility to noise and distortion, and more rapid processing. The disadvantage is that details smaller than the digitisation interval are either lost or only recovered slowly and with more difficulty.

In speech, two formants in the harmonic spectrum are used to categorise the vowels. The rapid variation of formants during transients also contributes to the recognition of consonants, as do the frequencies of the broad spectral bands, and timing cues. Together, these parameters are used to convey phonemes, and thus code the 'textual information'—the information retained in a transcription.

In languages without lexical tone, the fundamental frequency is both used and perceived as a continuous variable. It carries almost no information in the written text. Its absolute value and variation carry other information, but this information is not usually transcribed, is less explicit, and has rather less unanimously agreed meaning. Similar observations may be made about the duration and time separation of phonemes: with the exception of a small number of vowels distinguished by being long and short, timing information or rhythm convey little explicit information. They may be varied considerably with little effect on the transcription, with the possible exception of a small number of punctuation data. The prosody (pitch and rhythm) depends upon or may be interpreted as indicating the emotional state of and other information about the speaker. Thus, prosody overlaps very largely with what we have called the performance information: the information, usually not notated, that would be input *ad libitum* by the speaker of a written text.

Music uses these acoustical features quite differently. The spectral envelope, the formants, the frequency bands of broad band signal and the spectrum and envelope of the starting and finishing transients are all, in a musical context, related to timbre. They are not usually categorised: they correspond to analog variables. The acoustical features that encode the text in speech tell us in music what instrument is playing and how it is being played. They do not tell us what tune is being played. The only thing that they tell us about the text is the name of the instrument at the left of the staff.

The pitch of subsequent elements in music, and their duration and spacing in time, convey the melody and allow us to identify the music being played. These parameters encode the data carried in a transcription. Data used to convey performance information in speech are used to convey the 'text' of music.

Acoustical feature	**Music**	**Speech**
Fundamental frequency (when quasi periodic)	*pitch component of melody* categorised notated precision possible	*pitch component of prosody* not categorised not notated variability common
Temporal regularities and quantisation on a longer time scale	*rhythmic component of melody* categorised notated precision possible	*rhythmic component of prosody* not categorised not notated variability common
Short silences	*articulation* sometimes notated	*parts of plosive phonemes* implicitly notated
Steady formants	*components of instrumental timbre* not notated not categorised	*components of sustained phonemes* notated categorised
Varying formants	*not widely used* —	*components of plosive phonemes* categorised notated
Transient spectral details	*components of timbre* not categorised sometimes notated	*components of consonants* categorised notated

Table 1: Some acoustical features of music and speech signals.

4. ANALYSIS AND DECODING

Accurate communication and interpretation of detailed information using sound is a complicated process, as researchers in automatic speech recognition (and text-to-speech) attest. It is especially difficult when there is even a modest level of background noise. How did our species learn to do it in so short a time? How do individuals learn it in a much shorter time? Why is it so easy for us? Psychologists of perception have studied the perception of sound in great detail, but here I make only some general observations.

ISBN 1 876346 39 6

To track a single voice against a background signal, we use a range of cues. A set of harmonically related spectral components are perceived together, and their collective variation with time (melody, prosody or vibrato in the case of a singer) are used to identify an individual voice. This is a serious problem for speech and music recognition software. But not for us. Why not?

Even without noise, it is a complicated task to extract, from an acoustic signal the spectrum, envelope and pitch that carry information in both speech and music. In some cases, however, it is easier in music. A single line of melody, played by a single instrument or sung without words, has frequencies that remain stable for the duration of a note, whereas in speech the pitch usually changes continuously. The spectrum of this melodic sound changes less, and in a more regular way, than does the spectrum of speech. Further, many features of the timbre are thus held constant, or at least repeated, throughout the performance. Finally, melodies have more regular rhythms than speech. Teachers will recognise this as an example of reductionist strategy: begin with simple cases, holding most variables constant and varying others one or a few at a time. Is it possible that in singing to babies we are teaching them how to listen, preparing the skills necessary to understand speech? To what extent are the skills related [8]? If the use of vocal sounds has in the past enhanced the chances of survival or mating, then a propensity to use music in this way, whether transferred genetically or culturally between generations, may have been selected.

Singing might thus be compared to a game. Games are often analysed as models of 'serious' behaviour, and they can teach generally useful mental and physical skills. Physical games teach reflexes, co-ordination and muscular strength which may confer survival advantages. Intellectual and socialising games may also promote skills that confer survival or mating advantages. This is widely believed to be how the propensity to play games and the tendency to enjoy the exercise of useful physical and analytical skills are selected, whether genetically, culturally or both.

5. THE SOUND ANALYSIS GAME

To the extent that sound analysis is a 'game', its 'rules' are unwritten, complicated and variable. Comparing music to games does not trivialise music. Games may be quite complex, and humans may take them very seriously: cricket and chess are obvious examples. Our enjoyment of analytical exercise requires successively more complicated games as our analytical capacities develop. Perhaps our biological and cultural evolutions have left us with an innate enjoyment of neurological exercise and challenges, including this way of exercising our audition.

The subtle variations in timing, articulation and pitch used to carry expressive information in both speech and music, and the similarities in their uses, have been well studied (eg. [10]). However, to end this paper (and to open the discussion), I invite the reader to wonder whether the partial complementarity of the coding between music and speech may have something to do with the emotional power of music. Is it also possible that the coding of the textual information of music with the acoustical features used to convey affective, non-textual information in speech has something to do with the attraction of music, with the emotional potency and aesthetic appeal of this peculiarly coded, abstract method of communication?

6. REFERENCES

1. Copland, A. "What to listen for in music". New American Library, NY, 1967.
2. N.H. Fletcher and T.D. Rossing "The Physics of Musical Instruments". New York: Springer-Verlag, 1998.
3. Potter, R.K., Kopp, G.A. and Green, H.C. "Visible Speech", van Nostrand, New York, 1947.
4. Clark, J. and Yallop, C. "An Introduction to Phonetics and Phonology". Blackwell, Wiltshire, 1990.
5. Locke, S. and Kellar, L. "Categorical perception in a non-linguistic mode" *Cortex*, 9, 355-369, 1973.
6. Harnad, S. (ed.) "Categorical Perception: The Groundwork of Cognition", Cambridge University, 1987.
7. Jackendoff, L. "A comparison of rhythmic structures in music and language", in "Phonetics and Phonology. Rhythm and Meter" (Vol 1), Kiparsky, P. and Youmans, G., eds. Academic, San Diego, 1989.
8. Gérard, C and Auxiette, C. "The processing of musical prosody by musical and nonmusical children", *Music perception*, 10, 93-126, 1992.
9. Dowling, W.J. and Harwood, D.L. "Music Cognition", Academic, New York, 1986.
10. Banse, R. and Scherer, K.R. "Acoustic profiles in vocal emotion and expression" *J. Personality and Social Psychology*, 70, 614-636.

ISBN 1 876346 39 6