12
Music and the phonological principle: Remarks from the phoneticians's bench

J. B. Pierrehumbert

1. SOUND STRUCTURE IN LANGUAGE

1.1 *Phonology and its relation to syntax*

Language does not use an arbitrary collection of noises to convey different meanings. Instead, a small number of sound elements, which are themselves meaningless, are put together in various combinations to make words. In classical structuralist linguistics, this characteristic of language is referred to as "the phonemic principle". In subsequent work, the phoneme has given way to distinctive features and autosegmental tiers, but the idea is preserved that meaningful forms are built up by regularly combining a small number of meaningless sound elements. Let us therefore speak of "the phonological principle." Because of this fact about language, it is reasonable to develop a theory of phonology as such.

One of the main jobs of the phonology is to create representations which can serve as input to the phonetics. Specifically, the inventory of elements refers to dimensions of articulatory contrast, and the grammar which combines these elements creates pronounceable sequences by controlling their grouping, prominence, and coordination. Modern phonological theory makes detailed proposals about how the job of creating pronounceable representations is carried out. Some examples will serve to indicate that both structure and content are manipulated:

1) When words are combined into phrases, phonological rules supply a metrical structure for the phrase, building on the metrical structures for the words.

2) It is common for the result of concatenating two morphemes to be unpronounceable, according to the lights of the language in question. In this case, the phonology makes necessary repairs. For example, when the English plural suffix /s/ is added to a word like "fox", an epenthetic vowel is inserted so that the result conforms to principles of syllable structure.

3) In Lexical Phonology, underspecification theory accounts for segmental features which are specific (in the sense of being produced in some particular way)

but nondistinctive (in the sense of failing to distinguish between words). For example, the place of articulation of nasal consonants may be nondistinctive in a language (because nasals assimilate to a following stop); but it still must be specified (because a nasal consonant without a place does not make phonetic sense). Kiparsky (1985) and Ito (1988) provide detailed proposals for handling such cases.

A consensus has emerged in generative linguistics that phonological representations include a hierarchical structure, and that this structure is distinct from the syntactic structure which supports semantic interpretation. Thus we may view language as having two grammars. The phonological grammar specifies what is well formed *from the point of view of sound structure*. The syntax is responsible for the way that words are combined into sentences, including matters such as the antecedency of reflexive pronouns, and the order of verbs and their arguments.

The primary evidence for hierarchical structure in generative phonology has come from syllable structure, stress, and intonation. Kahn (1976) showed that many different processes of allophonic variation are conditioned by the same syllabic structure, so that strong generalizations are lost by permitting phonological rules to refer only to the segmental string, as in Chomsky and Halle (1968). The metrical structures proposed in Liberman and Prince (1977) supported a clear and elegant treatment of English word stress. These results inspired subsequent work on stress in other languages, leading to a universal parametric theory of stress rules (Hayes 1982, Halle and Vergnaud 1987). Extending the idea of metrical structure from the word level to the phrase level is supported by work on the assignment of phrasal intonation (Liberman 1975, Pierrehumbert 1980) as well as by work on phrase-level control of allophony (for example, Pierrehumbert and Talkin in press).

A distinction between phonological structure and syntactic structure is supported by several types of data, apart from the obvious differences in descriptive vocabulary. First, the behaviour of clitics indicates that phonological and syntactic structure are not isomorphic. These are pronouns, prepositions, auxiliaries and conjunctions which are treated phonologically as part of a neighboring word even though they stand as separate words syntactically, and may even be attached very high up in the syntactic tree. A second type of nonisomorphism is found when intonation phrases correspond to syntactic nonconstituents, as in [1] (where the diacritic % is used to indicate an intonational phrase boundary).

[1] I've never heard of a piece of legislation % which was more contrary to the public consensus % as overwhelmingly expressed in the last election.

Specific proposals for handling such cases are developed in Selkirk (1980, 1984) and Hirst (1987). Although they differ in some respects, they share the idea that prosodic phrasing is set up as a function of the syntactic and/or semantic structure. Both thus distinguish between the prosodic structure itself (in which syntactic and semantic information is not directly indicated) and the reasons for the prosodic structure (which are syntactic or semantic).

Third, the syntax underdetermines the intonational phrasing, so that the same syntactic structure can often be produced with several different intonational phrasings. Fourth, information which figures centrally in syntactic and semantic structure is lost to the phonetics. For example, there are many syntactic and semantic reasons to use an intonation break, including setting off a parenthetical, disambiguating the scope of an adverb, or providing a nuclear stress for more than one focus. However, the phonological and phonetic consequences are all the same. A similar conflation of syntactic and semantic information leads to a bewildering proliferation of excuses for the use of a pitch accent. Clearly, there would be a substantial loss of generality in a theory which provided a list of syntactic and semantic information which was unavailable to the phonetics. Instead, phonology undertakes to provide a positive specification of what *is* available. That is, we suppose that the rather meager resources of phonology mediate between syntax/semantics and phonetics.

Some particular aspects of the meagerness of phonology will be important for the discussion of musical performance below. First, phonological structures as presently understood are hierarchical but not recursive. That is, the phonological units of syllables are organized into feet, feet into phonological words, and words into phrases (Selkirk 1980, Selkirk 1984); however, we do not find rules in the phonological grammar such as [2], in which the same symbol appears both on the left and the right of a rule. Such rules are found in the syntax, as suggested in [3] .

[2] Syllable --> C V Syllable (impossible rule)

[3] NP --> NP S (expansion of NP with a relative clause)

(Recursion during the course of word derivation is posited in Lexical Phonology but it does not result in recursion in the surface representations). The rules for intonational phrasing developed in Hirst (1987) may be specifically interpreted as mapping the recursive structures of syntax into the relatively flat structures of phonology.

Second, phonological structures do not function meta-linguistically. That is, although phonological elements are involved in similarity relationships, they do not actually refer to other phonological elements (but only perhaps to the sounds used to realize them). Semantic and pragmatic theory, in contrast, do need to cover the meta-

linguistic function of language. It is needed not only to describe instances of quotation, but also the behaviour of discourse markers such as "now" and "then" which treat the text on a par with the real world context (e.g. Schiffrin, 1990).

Third, phonology has a weaker ability than syntax to define long distance dependencies. Specifically, metrical and autosegmental phonology provides a limited ability to describe nonlocal relationships by defining representations in which adjacency is defined over groups and melodic tiers, rather than over phonemes. They do not supply the formal power used in syntax and semantics to describe phenomena like relative clause formation or quantifier scope. The rules which access phonological representations are required to be local. This is true not only of phonological rules per se (Ito 1988), but also of the the rules which relate phonological elements to quantitative aspects of pronunciation (Pierrehumbert 1980, Liberman and Pierrehumbert 1984, Pierrehumbert and Beckman 1988).

1.2 Experimental support for hierarchical structures in phonology

Experiments on speech production support the existence of hierarchical structures in phonology by showing that the pronunciation of individual phonological elements depends not only on their identity, but also on their prosodic position. Specifically, pronunciation distinguishes elements which are at a prosodic boundary or the strongest in their group. The groups established by the prosody can also serve as a vehicle for expressive parameters.

Pierrehumbert and Talkin (in press) examined the pronunciation of the phoneme /h/ in various prosodic contexts. Their experiment manipulated the location of /h/ with respect to word and intonation phrase boundaries and word and phrasal stress. Acoustic indices of the degree of abduction of the vocal folds were developed. The behaviour of these indices showed that the pronunciation of the /h/ was determined by the interaction of word-level and phrase-level prosody. Specifically, all /h/s were affected by being in accented position (being in a rhythmically strong position in the phrase), but the effect differed depending on the position of /h/ with respect to word stress and boundaries. The study also found that /h/ was both longer and more strongly articulated when it followed an intonational boundary, even when there was no pause at the boundary.

Liberman and Pierrehumbert (1984) asked subjects to produce lists of berry names in three overall voice levels with a downstepping intonation pattern. In this pattern, each pitch accent is produced lower than the preceding one, so that a descending staircase results. The phonological treatment of Pierrehumbert (1980) describes the patterns in this experiment as involving a series of bitonal pitch accents H+L (High+Low), each of which triggers a lowering of fundamental frequency.

Figure 1 shows the results of this experiment for one subject; other subjects were similar.

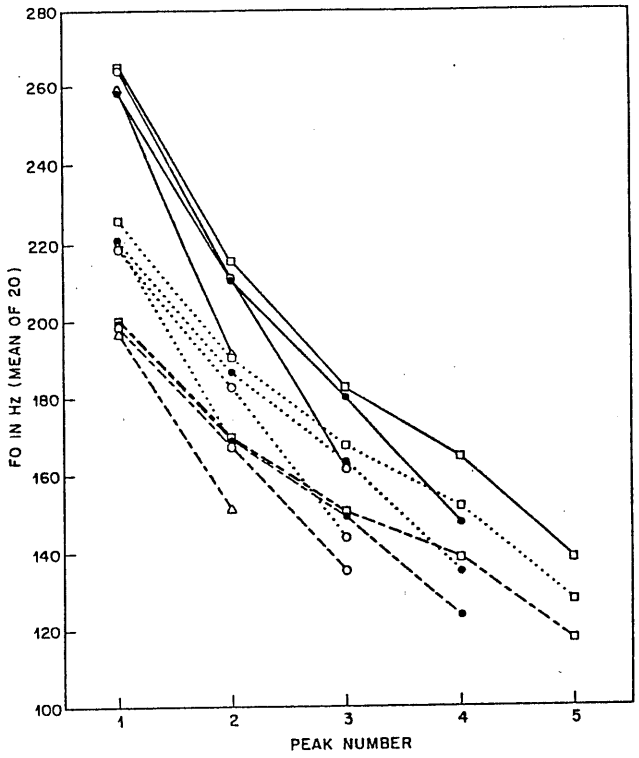
Four points may be drawn. The first concerns the downstep rule itself, which illustrates the fact that the pronunciation of a tone can depend on its tonal context as well as its structural position. Examples in which the production of a phoneme depends on its phonemic context are of course abundant, including all cases of assimilation and coarticulation. Such processes all access the context locally, in the sense that they refer only to elements in a small neighborhood; there is still controversy about just how small.

Second, we observe that the dictated pitch range affected the f0 values throughout the list. This means that the pronunciation of each individual pitch accent depends not only on what it is, but also on what the current pitch range is. This illustrates one function of nodes in phonological structure, which is to serve as a vehicle for expressive or stylistic parameters. Pierrehumbert and Beckman (1988) present results for Japanese indicating that pitch range is not just assigned to an utterance as a whole, but rather is hierarchically assigned to phonological constituents.

Third, in every list regardless of length, the f0 value for the last item is lowered. This final lowering, which was also observed for another intonation pattern discussed in the same paper, illustrates the relevance of structural position.

Fourth, nonfinal values are remarkably similar regardless of the number of upcoming elements. This result is somewhat surprising, since the experimental paradigm gave subjects maximal opportunity for preplanning, and they might have been expected to take smaller steps when more steps were to be accommodated in the pitch range. It supports the idea that phonetic implementation rules refer locally to properties of a hierarchical structure; the total number of terminal elements cannot be defined as a local property of such a structure.

A very important study by Gee and Grosjean (1983) examined pause durations in production of complex sentences. They compared models which predict pause durations as a function of boundary strength in the syntax with a prosodically based model which they developed, building specifically on Selkirk's work (1980, 1984). The syntax-based models (including in particular the model of Cooper and Paccia-Cooper, 1980) had some success in predicting pauses; this is not surprising in view of the semi-regular relationship between prosody and syntax and semantics, as discussed above. However, the prosodic model performed much better. It both explained more of the variance overall, and showed more freedom from systematic errors. In addition, it appeared to be more psychologically plausible since the units it established could be determined left to right, rather than by a global computation over the syntactic structure.



Downstep data (3 pitch ranges, 4 lengths) for subject DWS

2. MUSIC AND ITS RELATIONSHIP TO LANGUAGE

There is a substantial literature comparing music to language. In this literature, comparisons are made to many different levels of description of language. Separating the levels is difficult, because as many authors note it is hard to separate judgments of what is well-formed musically from judgments of what is musically meaningful. We may attribute this difficulty to the fact that music, unlike ordinary

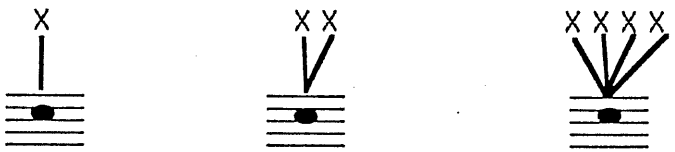
language, does not have an arbitrary association between form and meaning. Quite the contrary: The meaning of a musical fragment appears to be closely tied up with its key, melody, and rhythm, and so on, and meta-linguistic use of music appears to necessarily involve repetition. However, data on well-formedness is not the only kind that can be brought to bear on representational issues. As shown above, ordinary language displays a convergence between evidence from phonological well-formedness and evidence concerning the representations accessed during production. When we examine the results of experiments on musical performance and proposed rule systems for music synthesis, we find that it is possible to entertain the hypothesis that music has a well-defined level of phonological structure, which is broadly similar to the phonological structure of language. This is the case because the performance principles are found to depend on structures which are hierarchical but not recursive, and to access these structures in accordance with the locality principles of phonology. Although work in musical analysis and in algorithms for generating tunes manipulate properties which are outside of the purview of phonology, there is at present no evidence that these properties directly control performance.

2.1 Music and phonology

In order to clarify the comparison between music and language, let us begin by identifying some aspects of music which resemble phonology.

Music resembles phonology in grammaticalizing a temporal and acoustic space which is in principle continuously variable. The grammar involves both paradigmatic relations (covering the inventory of elements from which a choice can be made at each point) and syntagmatic relations (covering the types of grouping and prominence which can organize a sequence of elements). In order to maximize the similarity between music and phonology, let us suppose that note duration is treated structurally along the lines laid out in autosegmental phonology. Taking x to represent the shortest metrically relevant unit, we would then obtain representations like the following:

[4]



Having made this move, the metrical levels of music (such as the half-bar, the bar and the phrase) are then analogous to the levels of metrical phonology, with beats corresponding to stresses. This resemblance is noted in Lehrdahl and Jackendoff (1983).

Next to these broad similarities, a number of specific differences between the phonological level of music and that of language may be suggested, and more will no doubt be discovered. First, music manipulates more distinctive quantities than language. Although many languages have a two-way length distinction, the only serious candidate for a three-way distinction is argued in Prince (1980) to be reduceable to two. Rules of tone spreading and vowel harmony can give rise to representations in which a single element is linked to many timing slots; however, from a linguistic point of view, these cases involve indefinitely many slots, rather than a distinctive large number of slots. Second, extrametricality is more extensive in music than in language (Lehrdal and Jackendoff, 1983). Third, in music both melody and quantity are loosely related to structure (with the strictness of the relationship apparently varying by genre); in language, this relationship is tightly constrained by rules of syllable structure and stress. Such differences do not affect the broad analogy which is being drawn here and which is indeed intended to bring them out. They do, of course, give rise to reflections about possible differences in cognitive and social status between music and language.

Is music pure phonology? The essential incompleteness of a phonological treatment of music is shown by examples which patently involve a level of meaning. Musical compositions can deliberately quote other works; although quotation may be as uncommon in music as in ordinary language, it makes the same case for a semantic level by being metalinguistic. Sloboda and Parker's (1985) analysis of tune recall relies on the claim of Meyer (1978) that melodies "imply" chord sequences. Since semantic theory reserves the word "imply" for truth-conditional relations, one might propose substituting "implicate", also a concept in semantics. The iconic or descriptive use of music is well known; insofar as it involves learned conventions rather than direct imitation, it also motivates representation of musical meaning. Because of such facts, musical meaning in some form is universally acknowledged in the literature.

The status of musical syntax (as opposed to musical phonology) is a more delicate matter. On the one hand, it is hard not to conflate all structural aspects of music. For example, Sloboda (1985) distinguishes musical phonology from musical syntax, but views phonology as dealing only with the inventory of categories. On the other hand, separation of syntax from semantics is even more problematic for music than for natural language, as noted in Baker (1989). However, algorithms for generating compositions raise the prospect of a distinct musical syntax, at least when

viewed from a certain theoretical prospective. Sundberg and Lindblom (1976) pioneered this area with their algorithm for composing nursery tunes in the style of Tegner. Subsequently, Steedman wrote a grammar for chord progressions in the 12 bar blues (Steedman, 1984) and Baker (1989) also developed a grammar in order to specify the well-formedness conditions in his parser. Both Steedman (1984) and Baker (1989) use context free phrase structure grammars and specifically exploit the recursion available in such grammars. For example, Steedman says that his rule 3, which extends an authentic cadence backwards, "can quite correctly apply recursively to its own results" and that such application was in fact typical after World War II. Since phonology as now understood lacks recursive structures, such proposals point to a distinct level of syntactic representation for music.

1.2 Musical structure and performance

Results of experiments on musical performance show strong parallels between results on phonetic consequences of prosody. The performance of individual notes depends on their melodic context and structural position. As in speech, we find effects of being at a boundary, effects of being the most prominent element in a group, and effects which are spread over an entire group. Levels of grouping which are specifically supported are the half-bar, the bar, and the phrase. There are no results which suggest that recursive structures of the sort generated by grammars of chord progression bear directly on production. This is not because negative results exist, but rather the relevant experiments have not been done.

The tonal context plays an important part in the rules for computer simulation of expressive performance presented in Sundberg et al. (1983). These rules were based on informal impressions of performance by the authors. It is interesting to note that all rules make local use of the context; that is, they refer to neighboring elements of the note sequence. Sundberg et al. (1989) present data which also point to the importance of the immediate tonal context. In an experiment in which string players performed sequences of two notes, systematic deviations from equal-tempered tuning of the second note were found. The deviations exhibited a linear relationship to the "melodic charge", an objective measure related to the unexpectedness of the note. An alternative interpretation of the same experimental results would take the melodic charge to be a relationship between the current note and the key, viewed as a property of the phrase. This interpretation also falls within the formal power of phonology, as seen above in the discussion of effects of phrasal pitch range in speech.

The relevance of metrical prominence to performance is illustrated in a study by Sloboda (1983). He reports an experiment in which pianists performed the same note sequence with the bar lines and beams shifted to different locations. Because the sequences were embedded in context, no performer noticed that they were

identical. Notes at the half bar level were found to be marked by greater intensity, a more legato touch, and/or a delay in the onset of the following note. Since the half bar level is not notated in the score, the results provide evidence that the performer actively constructs a representation of the grouping and prominence.

The importance of the bar level is shown by experiments on piano playing reported in Shaffer (1981) and Palmer (1989). Both found that the asynchrony used to mark voicing is greatest at the first beat. Shaffer (1981) also reports results on the covariance of note durations which suggest that the bar acts as a unit of tempo; shifts in overall tempo take place between bars rather than within bars. This suggestion parallels the finding in speech that prosodic units act as vehicles for pitch range. It would be interesting to examine the prosodic domains for other stylistic parameters in the performance of music as well as in speech.

Studies of timing show that duration is used to mark phrase boundaries, just as duration is used to mark phrasing in speech. A number of results further show that the phrasing is constructed by the performer from the score, and that he has some discretion in doing this just as a reader can exercise choice in assigning phonological phrasing to a text. Palmer (1989) shows that phrasing as objectively indicated by rubato is the same as what performers mark on the score; the performers' own markings predict the rubato better than the (often different) analyses of other performers. A further study reported in Palmer (forthcoming) shows that expressive timing and the pattern of performance errors reflect the same phrasing assignment.

Shaffer's (1981) results, which are more extensive than there is space to summarize here, broadly support hierarchical control of timing, a concept pursued in Shaffer and Todd (1987). Shaffer and Todd themselves describe their experimental results as indicating "recursive" phrase final lengthening, and this description is worth evaluating carefully in view of its relevance to the comparison between the structures they propose and those of phonology and syntax. Both the data they present and their discussion of it suggest a prosodic hierarchy along the lines proposed for natural language, with lengthening reflecting the depth of the boundary in the hierarchy. Their hierarchy includes beats, measures, phrases, and even sections (although the need for the section as a prosodic unit is not defended in detail). Todd's (1989) discussion of these data and their implications for psychology and computational modelling specifically draws attention to the issue of placing realistic limits on access to global information during performance. He critiques his own earlier timing model, based on the time span reductions of Lerdahl and Jackendoff (1983), as leading to "more degrees of boundary strength and therefore degrees of relative slowing than can be discerned from the data", and argues instead for a model with a level-ordered hierarchy. I conclude that the timing rules need not manipulate recursive structures as generated by the Steedman and Baker grammars. Experiments similar to those of Gee and Grosjean (1983), specifically comparing syntactic and phonological structures as predictors of time patterns, appear not to

have been carried out for music. For example, it would be possible to collect data on performances of the 12-bar blues, in order to assess whether the timing is directly predicted by the depth of embedding as established by Steedman's grammar for chord progressions. Or, the implications for performance of Lerdahl and Jackendoff's level of prolongational reduction (evidently a structural semantic representation) could be explored. The hypothesis presented here would predict the timing would instead reflect the consequences of mapping into a flatter prosodic representation.

All of the experiments just summarized provide support for a hierarchical structure by showing that the performance of notes differs according to structural position. As discussed in section 1, a hierarchical representation also carries with it the implication that certain aspects of the context should fail to affect performance. Although Todd (1989) makes note of this point, the only experimental result I have found which specifically addresses it is reported in Sloboda (1985). Sloboda investigated the eye-hand span in sight-reading by recording how far a pianist could continue after the score was unexpectedly removed. An average span of about seven notes was established, clearly shorter than the typical phrase. However, the length of the span was affected by the phrasing. The span could be somewhat longer or shorter than average so as to end at a musically defined phrase boundary. Further experiments establishing differential access to context during performance could help to clarify the representation of musical structure.

3. CONCLUSION

Experimental data on speech production and on music performance show broad similarities. In both cases, the production system appears to have local access to a prosodic structure which is hierarchical but not recursive. Thus the evidence to date supports a strong analogy between the performance structures of music and the structures of phonology, permitting the conjecture that musical performance like speech is phonologically mediated. Taking seriously the formal limitations of phonology might make it possible to evaluate this conjecture.

4. REFERENCES

Baker (1989). An Artificial intelligence approach to musical grouping analysis. Contemporary Music Review 3(1), 43-68.

Cooper and Paccia-Cooper (1980). *Syntax and speech*. Harvard Univ. Press, Cambridge MA.

Gabrielsson, A., ed. (1987).*Action and perception in rhythm and music*, Royal Swedish Academy of Music, Stockholm.

Gee and Grosjean (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology 15*, 411-458.

Halle, M. and J.-R. Vergnaud (1987) *An Essay on Stress*. MIT Press, Cambridge.

Hayes, B. (1980) *A Metrical Theory of Stress Rules*, Ph.D Dissertation, MIT. Distributed by Indiana University Linguistics Club, Bloomington.

Hayes, B. (1982) "Extrametricality and English Stress," *Linguistic Inquiry 13*, 227-76.

Hirst. D. (1987) La description linguistique des systemes prosodiques: une approche cognitive. These de Doctorat d'Etat, Universite de Provence.

Howell, P. , I. Cross and R. West , eds. (1985). Musical Structure and Cognition. Academic Press, London.

Ito, J. (1988) Syllable Theory in Prosodic Phonology, Ph.D dissertation, U. Mass Amherst. Published by Garland Publishing, New York.

Kahn, D. (1976) Syllable Based Generalizations in Generative Phonology. Ph.D dissertation, MIT. Published 1980 by Garland, New York.

Kiparsky, P. (1985) "Some Consequences of Lexical Phonology," Phonology Yearbook 2, 85-138.

Lerdahl, F. and R. Jackendoff (1983) A Generative Theory of Tonal Music. MIT Press, Cambridge MA.

Liberman, M. Y. (1975) The Intonational System of English. Ph.D dissertation, MIT. Published 1979 by Garland, New York.

Liberman, M.Y., and J. Pierrehumbert (1984) Intonational Invariance under Changes in Pitch Range and Length. In Aronoff, M., Oehrle, R.T. (eds.), *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, MIT Press, Cambridge, 157-233.

Liberman, M.Y. and A. Prince (1977) On Stress and Linguistic Rhythm. *Linguistic Inquiry 8*, 249-336.

Meyer , L. B. (1973). *Explaining music*. University of California Press, Berkeley.

Palmer , C. (1989) Mapping Musical Thought to Musical Performance. *Journal of Experimental Psychology: Human Perception and Performance 15 (12)*, 331-346.

Palmer. C. (in press) The Role of Interpretive Preferences in Music Performance. Cognitive Bases of Musical Communication.

Pierrehumbert, J., (1980) The Phonology and Phonetics of English Intonation, MIT Ph.D dissertation. Available from Indiana University Linguistics Club, Bloomington, Indiana.

Pierrehumbert, J. and M. Beckman, (1988), *Japanese Tone Structure, Linguistic Inquiry Monograph Series 15*, MIT Press, Cambridge.

Pierrehumbert, J. and D. Talkin (in press), "Lenition of /h/ and glottal stop," in Ladd and Doherty, (eds.) *Papers in Laboratory Phonology II*, Cambridge University Press, London.

Prince, A. (1980) A metrical theory for Estonian quantity. *Linguistic Inquiry 11*, 511-62.

Schiffrin , D. (1990) Between Text and Context: the Meaning and Use of "Then". *Text 10(3)*.

Selkirk, E. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry 11*, 563-605.

Selkirk, E. (1984). *Phonology and Syntax: the relation between Sound and Structure*. MIT Press, Cambridge, MA.

Shaffer, L. H. (1981) Performance of Chopin, Bach, and Bartok: studies in motor programming. *Cognitive Psychology 13*, 326-76.

Shaffer. L. H. and N. P. Todd (1987) The interpretive component in musical performance. in Gabrielsson, ed., 139-152.

Sloboda, J. (1983). The communication of musical meter in piano performance. Q. Jl. Exp. Psychol. 35, 377-396.

Sloboda, J. (1985). *The Musical Mind: The cognitive psychology of music*. Clarendon Press, Oxford.

Sloboda, J. and D. Parker (1985). Immediate Recall of melodies. In Howell, Cross, and West (eds.), 143-167.

Steedman, M. (1984). A Generative Grammar for Jazz Chord Sequences. Music Perception 2(1), 53-78.

Sundberg, J., A. Askenfelt and L. Fryden (1983). Musical Performance: A Synthesis by Rule Approach. *Computer Music 7(1)*, 37-43.

Sundberg, J. and B. Lindblom (1976). Generative theories in language and music descriptions. *Cognition 4*, 99-1222.

Sundberg,J., Friberg and Fryden (1989). Rules for automated performance of ensemble music. *Contemporary Music Review 3(1)*, 89-111.

Todd, N. P. (1989). A Computational model of rubato. *Contemporary Music Review 3(1)*, 69-88.

13
Emotion expression in speech and music

K. R. Scherer

"Music is our oldest form of expression, older than language or art; it begins with the voice, and with our overwhelming need to reach out to others. In fact music is man far more than words, for words are abstract symbols which convey factual meaning. Music touches our feelings more deeply than words and makes us respond with our whole being."

This rather poetic quote from Yehudi Menuhin (Menuhin & Davis, 1979, p. 1) is intended as a reminder that voice and emotion have played a central role in the co-evolution of music and language and still represent vital and powerful elements in present day speech communication and music making. It is certainly true that in the course of human phylogenesis both music and language have evolved towards complex symbolic systems with components and rule structures that largely depend on the development of human cognitive processing capacities unprecedented in the course of evolution. The latter has given rise to the development of sophisticated instruments for the production of a virtually limitless variety of musical sounds and sound combinations together with the invention of elaborate formal prescriptions for composing music. As for language, complex syntactic and semantic rule systems for the representation of meaning and various non-vocal means for language production and transmission have evolved. During the past decades, most scientific analyses of language and music have focused on these formal systems, stressing competence aspects rather than performance. The pendulum of scientific fashion is about to swing back and interest in the voice as the primary human instrument for language and music production - speech and singing - is growing, particularly with respect to the voice as a medium of expression of emotion. I am convinced that the vocal expression of emotion constitutes a central element of the investigation of the interrelationships between music, language, speech, and brain. In this paper, I will advance some theoretical arguments to this effect followed up by a summary review of some of our own research efforts in this area.

The large majority of all animal vocalizations are affective in nature (which does not rule out that representational functions can be served at the same time; see Scherer, 1988). Ever since the pioneering work by Darwin (1872/1965), students of animal communication have demonstrated the important role of vocalization in the expression of affect (Marler & Tenaza, 1977; Morton, 1977; Scherer, 1985; Tembrock, 1975). In close parallel to animal affect vocalizations, we still find rudiments of non-linguistic human affect vocalizations, often referred to as "interjections", such as "aua", "ai", "oh", "ii", etc. These may have been more or less domesticated by a specific phonological system (see Scherer, 1977; Wundt, 1903). Affect vocalizations are the closest we can get to the pure biological expression of emotion and one of the most rudimentary forms of communication.

Given that the origin of language and of music are still shrouded in mystery, one might reasonably speculate that both proto-speech and proto-music might have used affect vocalizations as building blocks. Considering the way that human beings use affect vocalizations and vocal emblems (see Scherer, 1977) as a means of communication (i.e. in cases of lack of speech ability or language differences) and certain songs resemble conventionalized affect vocalizations (e.g. wailing patterns in mourning rituals), this may not seem too unreasonable an hypothesis. Helmholtz (1863/1954, p. 370-1) noted: "... a great part of the natural means of vocal expression may be reduced to such facts as the following: its rhythm and accentuation are an immediate expression of the rapidity or force of the corresponding psychical motives - all effort drives the voice up - a desire to make a pleasant impression on another mind leads to selecting a softer, pleasanter quality of tone - and so forth. An endeavour to imitate the involuntary modulations of the voice and make its recitation richer and more expressive, may therefore possibly have led our ancestors to the discovery of the first means of musical expression, just as the imitation of weeping, shouting, or sobbing, and other musical delineations may play a part in even cultivated music (as in operas)...".

It might well be, then, that the externalization of affect or emotion via vocalization is at the very basis of music and speech. As ethologists have shown (see Andrew, 1972; Leyhausen, 1967), expression and impression are closely linked. In the process of conventionalization and ritualization, expressive signals may be shaped by the constraints of transmission characteristics, limitations of sensory organs, or other factors (see also the discussion of push vs. pull effects, in Scherer & Kappas, 1988). The resulting flexibility of the communication code may have fostered the evolution of more abstract, symbolic language and music systems. This development is likely to have occurred in close conjunction with the evolution of the brain. Just as newer neocortical structures with highly cognitive modes of functioning have been superimposed on older "emotional" structures such as the limbic system, the evolution of human speech as a digital system of information encoding and transmission (and of musical scales and conventions for singing) has made use of the more primitive, analogue vocal affect signalling system as a carrier signal. In making use of vocalization, which continued to serve as a medium for emotion expression, as the production system for the highly formalized systems of language and music, the functions became by necessity strongly intermeshed. Thus, in speech, changes in fundamental frequency (F0) contours, formant structure, or characteristics of the glottal source spectrum, can, depending on the language and the context, serve to communicate phonological contrasts, syntactic choices, pragmatic meaning or emotional expression. Similarly, in music, melody, harmonic structure, or timing may reflect sophisticated constructions of the composer, depending on specific traditions of music, and may simultaneously communicate strong emotional moods (see Copeland, 1939/57; Meyer, 1956; Seashore, 1938/67). This fusion of two signal systems, which are quite different in function and in structure, into a single underlying production mechanism, vocalization, has proven to be singularly efficient - for the purpose of effective communication and from the point of view of evolutionary survival. It has also proven to be singularly messy and complicated for scientific analysis.

Linguists and phoneticians, with a few exceptions, have largely shunned the study of what has been called para- or extralinguistic communication. Even the study of prosodic, supra-segmental phenomena, which are of major importance for speech production and comprehension, has been greatly neglected until fairly recently - often with the justification that such "emotional" or "attitudinal" aspects of speech did not fall into the research domain of speech scientists (some notable exceptions notwithstanding, e.g. Bolinger, 1964; Crystal & Quirk, 1964; Williams & Stevens, 1972). Interestingly enough, it has been the practical necessity of teaching to speak a particular language to foreigners and to computers that has sparked major research efforts in this area, reflecting the fact that natural, acceptable speech depends on the production of appropriate prosodic features. Many of these are invariably affected by the speaker's attitude and emotional state. Prosody and other paralinguistic aspects of speech have been equally neglected by psychologists: Psycholinguists generally limit their interests to cognitive aspects of speech production and comprehension, and psychologists studying emotional expression have