Contemporary Music Review,
1989, Vol. 4, pp. 391–404
Photocopying permitted by license only

Speech and music performance: Parallels and contrasts

Rolf Carlson, Anders Friberg, Lars Frydén, Björn Granström & Johan Sundberg[1]

Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Box 70014, S-10044 Stockholm, Sweden

Speech and music performance are two important systems for interhuman communication by means of acoustic signals. These signals must be adapted to the human perceptual and cognitive systems. Hence a comparitive analysis of speech and music performances is likely to shed light on these systems, particularly regarding basic requirements for acoustic communication. Two computer programs are compared, one for text-to-speech conversion and one for note-to-tone conversion. Similarities are found in the need for placing emphasis on unexpected elements, for increasing the dissimilarities between different categories, and for flagging structural constituents. Similarities are also found in the code chosen for conveying this information, e.g. emphasis by lengthening and constituent marking by final lengthening.

KEY WORDS: Communication, duration, intonation, music performance, music synthesis, phrasing, prosody, speech synthesis.

Introduction

Much has been written and said about parallels between language and music regarding their structural aspects (Winograd, 1968; Bernstein, 1976; Sundberg & Lindblom, 1976; Lerdahl & Jackendoff, 1983). Striking similarities are generally found, such as a hierarchic structure with several levels. This hierarchical structure reflects the relations between different parts of the messsage. These relations have basic simple forms. With the help of variations and violations of this form, a speaker can focus on various important aspects of the message. The communicative success of these transformations is dependent on the listener's knowledge of the basic form and the speaker's presupposition about the listener's knowledge. In speech and music research, the relation between the speaker/performer and the listener at all structural levels is, or should be, of prime interest. In the present article we will concentrate on some interesting similarities and contrasts between language and music that seem to exist at the surface level.

Both language and music are realized in acoustic signals. Linguistics has introduced the term *speech* for the acoustic realization of language. In music science, no corresponding terminological distinction has been established. In this article we will use the term *music performance* for the acoustic realization of music. Thus, while the term language is equivalent to music, the term speech is equivalent to music performance.

One major problem both in speech and music performance research is that many different prosodic factors are mixed together as one single acoustic parameter. For instance, segmental inherent pitch, word tone, sentence type, lexical stress, emphasis etc. can all be signaled by one single parameter, such as the voice fundamental frequency. In the same way, the duration of speech sounds is affected by a variety of conditions including stress, position in the utterance, and local phonetic context. An extensive review of the factors that have been found to influence the duration of speech sounds are reported in a paper by Klatt (1976) and in special issues of *Phonetica* (1981, 1986). All of this is taken into account by the listener in the perceptual decoding process.

The same applies to music. There are many different reasons to lengthen or shorten a note beyond its nominal duration as specified in the score. Apparently, such perturbations of the nominal duration serve different purposes, e.g. emphasis, marking of phrase endings, and sharpening the contrast between categories. This results in rather complicated interactions making it hard to use conventional analytic methods. As a consequence, analysis-by-synthesis is a powerful tool in both speech and music performance research.

In recent years, the use of computers has led to great advances in both speech and music sciences. With the help of fast analysis tools, new knowledge has been gained and new models simulated. It is now possible to study the acoustic behavior of articulatory models or to compare duration models to natural recorded speech in data banks.

Speech recognition systems have attracted a lot of research money and some of it has been fruitfully invested in basic speech research. Text-to-speech programs have been productive in increasing the understanding of the speech communication process. With the help of models formulated as transformation rules, we have been able to test our current knowledge and to reject or accept ideas.

Similarly, our work with developing computer-generated music performances, by means of note-to-tone programs, has been revealing as to basic aspects of music communication. In this article, we will analyze some similarities that we have observed in our parallel working with text-to-speech and note-to-tone programs.

Text-to-speech and note-to-tone programs

We have previously reported on the long-term effort to develop high quality text-to-speech systems for several languages (Carlson & Granström 1975a; Carlson, Granström & Hunnicutt, 1982). The approach taken has been to formulate the process in a coherent framework. One criterion was that linguistics involved in creating, refining, and maintaining the text-to-speech software should be able to work with constructs and conventions familiar to them without necessarily mastering conventional computer programming. Consequently, *distinctive features* and *phonemes* are primes in our system. Also, the rule notation borrows heavily on that used in generative phonology, although it is expanded to easily handle continuous variables such as synthesizer parameters. This makes it possible to formulate pronunciation rules for coarticulation and utterance final lengthening in the same framework as the rules used for translating spelled text into phonetic transcription. Another important goal was to streamline the transfer

to a real-time system, which has the dual advantage of speeding up the testing of rules and of facilitating practical use in different applications.

Our current text-to-speech system consists of a structure of rule components and various lexica. The lexica have two important functions. The most important is to separate function words from content words. The function words, e.g., articles, prepositions, and pronouns, are mainly used to build up the grammatical structure of a language. Each language has a limited number of such words. These words are normally unstressed and often violate the pronunciation rules in a language. Content words follow pronunciation rules to a greater extent. The rules are written in the same formalism throughout the system and, in order to refer to different-level units, we attach appropriate features to our single stock of symbols. In this way everything from syntactic analysis to detailed sub-phonemic manipulations is handled in a coherent way.

The note-to-tone program is basically the same as the phonetic part of the text-to-speech program and was reformulated for musical purposes by coauthors Carlson and Granström (Sundberg, 1978; Sundberg, Askenfelt & Frydén, 1983). The modification regards the input format, being a transformation of the musical score, which can be executed automatically. Each note is defined as a sound possessing a pitch name (e.g. F sharp), an octave number and a duration, and each voice of a score is stored, represented in this way, as a separate input file.

Apart from this, information is also added regarding chords and phrase and sub-phrase boundaries. It would have been possible to write computer programs also performing chord and phrase analysis. However, this project aims at a description of music performance rather than music analysis. Therefore, it appeared preferable to do these analyses in the conventional way, i.e. using one's musical judgement.

The musical equivalents of the pronunciation rules in the text-to-speech program are the *performance rules.* These have been gradually developed over a long period of time. In this process, a professional music teacher and musician, Lars Frydén, has played a decisive role. He has "instructed" the computer on what to change in the performance in order to improve it musically. These instructions have been organized into a set of ordered, context-dependent rules, most of which have been tested in experiments with panels of expert listeners (Thompson, Friberg, Frydén & Sundberg, 1986).

In summary, parallel experiences have accumulated from working with speech synthesis (coauthors Carlson and Granström) and music performance synthesis (coauthors Friberg, Frydén and Sundberg) using similar tools. This allowed a more detailed comparison than is usually possible. Therefore, it seemed an interesting task to compare these experiences and to discuss the implications of similarities and dissimilarities.

An important difference in focus should be mentioned at this point. In the case of speech, the instrument producing the acoustic signal consists of the human voice organ. This is an exceedingly flexible instrument producing a highly variable acoustic signal. Consequently, synthesizing the output is a complicated task, and intelligibility rather than beauty has been the primary concern in this synthesis work.

Another difference between a normal text and music is also of general interest. It is relatively easy to generate the phonetic transcription from the spelled text.

In this transcription, only the segmental quality is marked. On the other hand, it is a major task to make some kind of syntactic and semantic interpretation of the text. Such information is in most cases needed in order to generate really convincing speech performance. In present text-to-speech systems, this problem is normally approached simply by identifying function words and perhaps accessing parts of speech (word class) information in a lexicon.

Normally, spelled text or phonetic transcription lack markings of phrases, emphasis, and most other indicators of higher-level relations. This has led current text-to-speech systems to concentrate on lower levels and primitive syntax analysis. Much less has been studied to formulate the acoustic correlates of larger syntactic structures. Even discourse structure, however, is known to affect, for example the acoustically realized duration and intonation (Lehiste, 1975).

The problems presenting themselves in attempts to synthesize music performance are, in part, different. By contrast to the vocal tract, most music instruments are built of rigid parts of wood or metal. This leads to an acoustically much more stable, less flexible signal, so that synthesizing the sheer instrument signals per se is by no means as complicated a task as in the case of the human voice. Therefore, in music performance synthesis, the main concern has been the expressive deviations by means of which the performer shows his musical interpretation of the piece. Fortunately for the musician, the music also contains more information about the composers' intentions at a higher level. This difference between speech and music performance synthesis work limits the possibilities of making exhaustive comparisons.

The human voice is undoubtedly a musical instrument when used for singing. Parallels between speech and singing are too obvious to be interesting and will mostly be disregarded in this paper.

Basic contrasts

Representation Levels

Both speech and music can be represented by means of graphical signs, viz., the orthography and the music score. However, in the case of speech, the phonetic transcription exists as an intermediate level of representation between orthography and speech. The relation between the orthography and the phonetic representation varies between languages; in some languages such as Finnish or Spanish, the relation is very straightforward making the intermediate level almost unnecessary. In other languages, like English, the relation is quite complicated and not entirely according to a set of conventions/rules. In this case, the speaker or synthesis program needs to rely on a phonetic transcription.

In a sense, an equivalent intermediate level of description exists also in music, viz., when the player or the composer complements the score by a great number of additional signs, such as dots, dashes, wedges, slurs, etc. This type of score has not formally been distinguished from the orthographic representation as clearly as in the case of speech. Also, this intermediate level seems more needed in speech. This can be concluded from the fact that the speech produced by the concatenation of sounds directly corresponding to the letters not only sounds extremely unnatural, but is also even practically impossible to understand; music produced from a nominal realization of the note signs, on the other hand, is still recognizable, even though very boring to listen to.

Quantization

Although both music and language can be represented by graphical signs that in some way can be regarded as symbols for the corresponding acoustic signals, the relationship between this graphical representation and the sound signals differs in one important respect.

In the music score, pitch and duration are represented by symbols according to a system of *quantized categories*. For instance, a quarter note (crotchet), is nominally twice as long as an eighth note (quaver), and a C is almost 6% lower in fundamental frequency than a C sharp.

In orthography, on the other hand, neither pitch, nor duration are specified in quantitative terms. In most languages it is rather the time derivative of pitch that is predictable from the orthography, such as in cases of question, quotation, accents, etc. or from the linguistic content, such as in cases of so-called focus. (Focus is a term used in linguistics to mark the part of the sentence which carries the most important information). In so-called tone languages, pitch is predictable within non-quantized categories, such as high, middle, low, rising, and falling. Duration can sometimes be predicted qualitatively from orthography, such as "long" and "short" vowels and consonants. However, prediction of phonetic transcription sometimes fails so that a lexicon is needed in order to arrive at correct word pronunciation.

Parallels

Role of the Author/Composer

The fact that we can formulate rule systems that generate intelligible speech and music performance of a decent musical quality apparently means that the acoustic realization is implied in the orthography and the music score. This suggests that the author limits the number of possible acoustic realizations of his text in a similar way as the composer limits the possible acoustic realizations of his score. This is probably an essential requirement on a useful symbol system, such as the orthography and the music score. This similarity indicates that, in this regard, similar processes underlie speech and music performances. However, the final acoustic realization is not of the same prime concern for authors as for composers. This is also why writing systems can have a more vague relation to the acoustic realization.

Stress and Emphasis

An important premise of the comparisons carried out in the present article is the difference between *emphasis* and *stress* in speech. While emphasis is content-dependent, the distinction between stressed and unstressed is a word-level phenomenon. This means that emphasis and stress exist at different levels in speech, stress being at a lower level. It seems that an equivalent distinction can be made in music, in that stress is a property that is dependent on the position in the bar, while emphasis is rather dependent on higher-level aspects of the musical structure.

Communicative Purposes

Predictability and emphasis

In both speech and music, a listener can predict most acoustic events due to his linguistic or musical experience and competence. For instance, we are very skilled in filling gaps in the acoustic information occurring because of interference with noise. Even if a door is banging in the middle of somebody's speech, we can mostly hear what the person is saying. It is often even hard to tell which speech sound was masked by the noise. Some classical perceptual studies have shown that the perceived location in time of a disturbing sound is judged to be a place in the conversation where it causes as little harm as possible, i.e., close to a syntactic break.

However, the meaning of an utterance does not always survive a door bang. In some places, the sentence is vulnerable and if information is lost in such places, the meaning of the sentence could not be restored. From this, we can conclude that predictability varies along a spoken utterance.

A similar reasoning seems applicable to music. Mostly we can complement a melodic line correctly, even if one note is missing. For instance, deleting a passing-note would be completely harmless. According to Sloboda (1977) the eyes of musicians tend to jump to beginnings and endings of phrases, thus apparently inferring the intermediate notes. On the other hand, there are also more important notes in a melody. In the second theme from the first movement of Schubert's *B minor symphony, D759*, there is a modulation from D major to B major, see Figure 1. The modulation is announced by a D sharp. If this note is cut out, the harmonic interpretation of the following notes will be affected. Thus, this D sharp seems to be a crucial note for the melody. It seems obvious that such important notes have a low predictability, and it can be assumed that predicability is dependent on the inverse of the information rate.

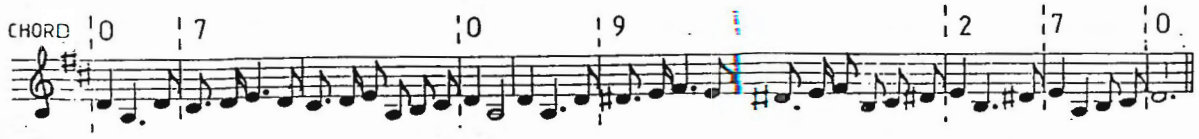

Figure 1 Second theme from first movement of Schubert's *Symphony in B minor, D 759*. The top line of numbers symbolizes the harmonies in terms of the distance, in semitones, between the root of the chord and the root of the tonic.

The parallel ocurrence of a time-varying predictability in both speech and music is by no means trivial. Its existence in both suggests that possibly it represents a way of meeting an essential limitation of the perceptual system. For instance, this system may be incapable of processing signals having an invariably high information rate.

Predictability of words has been studied by several authors. In a now classical study by Lieberman (1963), the relationship between context redundancy and keyword intelligibility was studied. It was shown that predictability had a strong effect on how clearly a word was pronounced. This experiment was later repeated by Hunnicutt (1985, 1987a). In a text-to-speech system, Coker, Umeda & Browman (1973) included factors such as word frequency and repetition of earlier mentioned words. These parameters added to the naturalness of the speech

quality. Similar ideas about predictability are included in modern communication aids for the handicapped. It can be shown that a prediction program working only on the surface structure can predict at least 50% of the typed letters in a running text (Hunnicutt, 1987b).

Predictability in speech is present at many different levels such as phoneme sequence, choice of word endings, and even syntatic constructs. At a structural level, we know that certain phrase or word combinations are very probable. It is even likely that certain word sequences are lexicalized just like certain single words, and that we perceive these phrases as single units.

The varying predictability in a sentence is significant to its acoustical realization, i.e., to speech. In order to make speech easy to understand and natural sounding, it is necesary to emphasize important, or less predictable words, and to deemphasize unimportant, predictable elements.

The varying predictability is significant also to the quality of music performance. In our rule system for music performance we have introduced a notion which we have called *melodic charge* in order to take into account the need for emphasizing certain notes and deemphasizing others. Melodic charge is defined with the aid of the circle of fifths, where the root of the prevailing chord has zero melodic charge, as illustrated in Figure 2. It is correlated with listeners' concepts of what is a good continuation of a scale, according to probe tone rating experiments carried out by Krumhansl & Kessler (1982). Also, according to Knopoff & Hutchinson (1983), it is positively correlated with the occurrence of the various scale tones in Schubert songs, so that, in some sense, it would reflect the negation of predictability, or *remarkableness* of the notes in a melody.

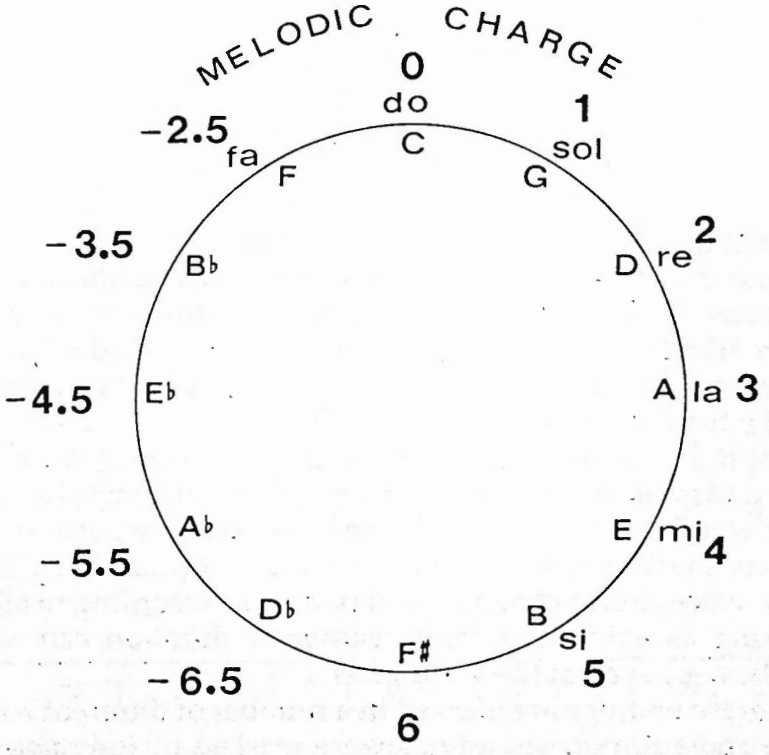

Figure 2 Definition of melodic charge by means of the circle of fifths.

The fact that the negation of predictability is marked by emphasis in both speech and music performance seems to suggest that adjusting the performance to the predictability is an important property of communication by acoustic signals.

Category separation

Some rules in the music performance program seem to serve the purpose of increasing the contrast between different categories. For example, minor seconds are played narrower than their nominal value, and major seconds are played wider. Also, short notes are shortened and long notes are lengthened. Another example is the *inégalles*, a playing manner from the Baroque era which is applied to long series of notes of the same nominal duration; in such series, duration is transferred from the unstressed to the stressed notes, the quantity transferred being slightly less than in the case of a punctuation. The result is that emphasis is laid on the differences between stressed and unstressed positions in the bar.

In speech, pitch and duration are specified in non-quantized categories as previously mentioned. Therefore, contrast sharpening is more difiicult to discern in speech than in music performance. Still, it is evident that categories exist also in speech. For example, there are long and short vowels, and the contrast between these categories is increased by formant frequency differences. Thus, a short Swedish /a/ has higher first and second formants than a long /a/. Word emphasis constitutes another example. In a study by Erikson & Rapp (described in Carlson, Erikson, Granström, Lindblom & Rapp, 1975), where emphasis was placed on different words in the same sentence, it could be seen that the syllables neighboring the emphasized word were pronounced with shorter durations and lesser fundamental frequency movements, while the duration of the emphasized word was increased and its fundamental frequency was quickly changing (see Figures 3 and 4). Finally, a classic observation from English is the marked difference between the length of a vowel before voiced and voiceless consonants. Phonetic contrasts are also exaggerated when there is a contextual need to distinguish two similar words, like Emigrant/Immigrant.

Constituent marking

Both speech and music have a hierarchical structure, as mentioned, and one of the obvious parallels is that phrases and other constituents in the hierarchical structure are marked in both. Moreover, the structure is evident from the graphic representation in both, at least for a human observer who is familiar with the language or the type of music.

In orthography, various signs such as periods, commas, semicolons, and spaces are used to mark the structure; but less explicit structural indicators are also used, such as word order and special words like conjunctions. In speech, these constituents are marked in many different ways, primarily by durational and modifications, voice source changes, and pausing. According to Klatt (1976) constituent marking as evidenced from segmental duration can be observed at different levels, e.g., even at the word level.

In music, phrase endings are marked in a number of different ways; in nursery tunes, phrase and sub-phrase endings were marked by long notes, and certain harmonic and melodic stereotypes (Lindblom & Sundberg, 1970).In the music performance program, phrase and sub-phrase endings are marked in the input notation. Then, the final notes of these two constituents are marked in the

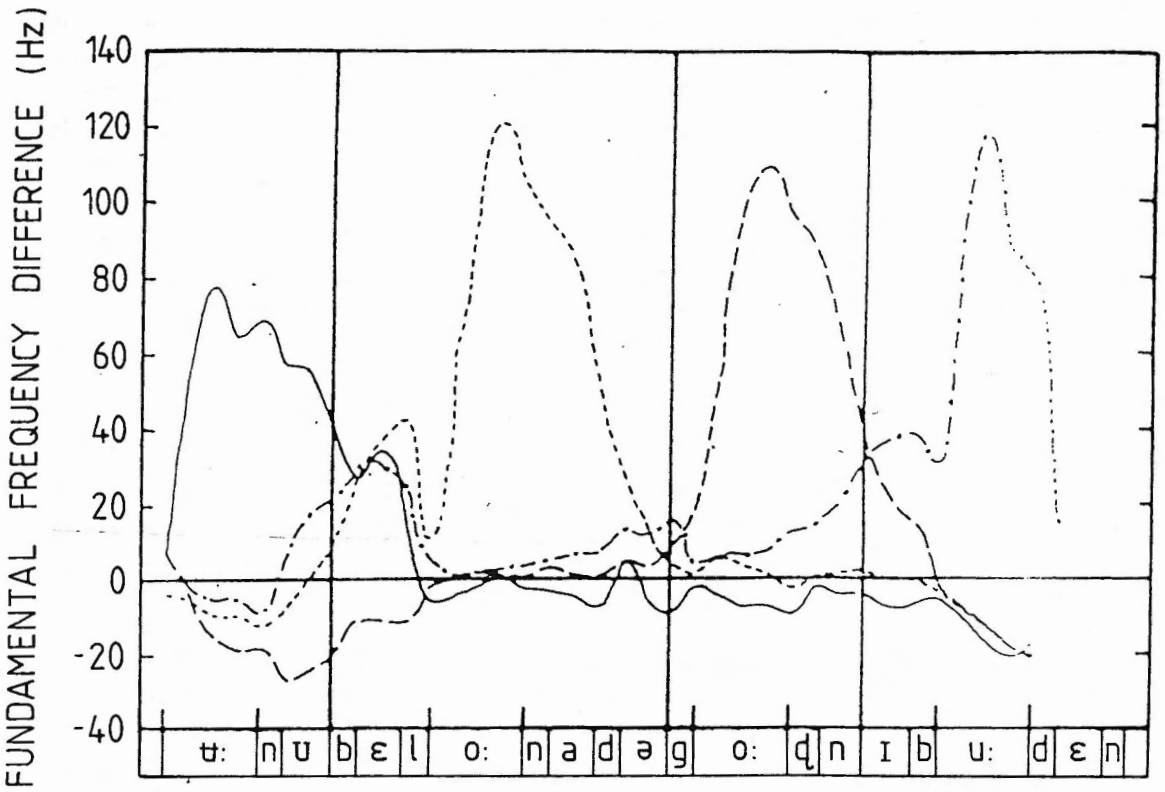

Figure 3 Fundamental frequency differences between emphatic productions and the averaged neutral production of the Swedish sentence "Uno belanåde gården i Boden", (Uno mortgaged the farm in Boden). The different curves pertain to emphasis on the four main words.

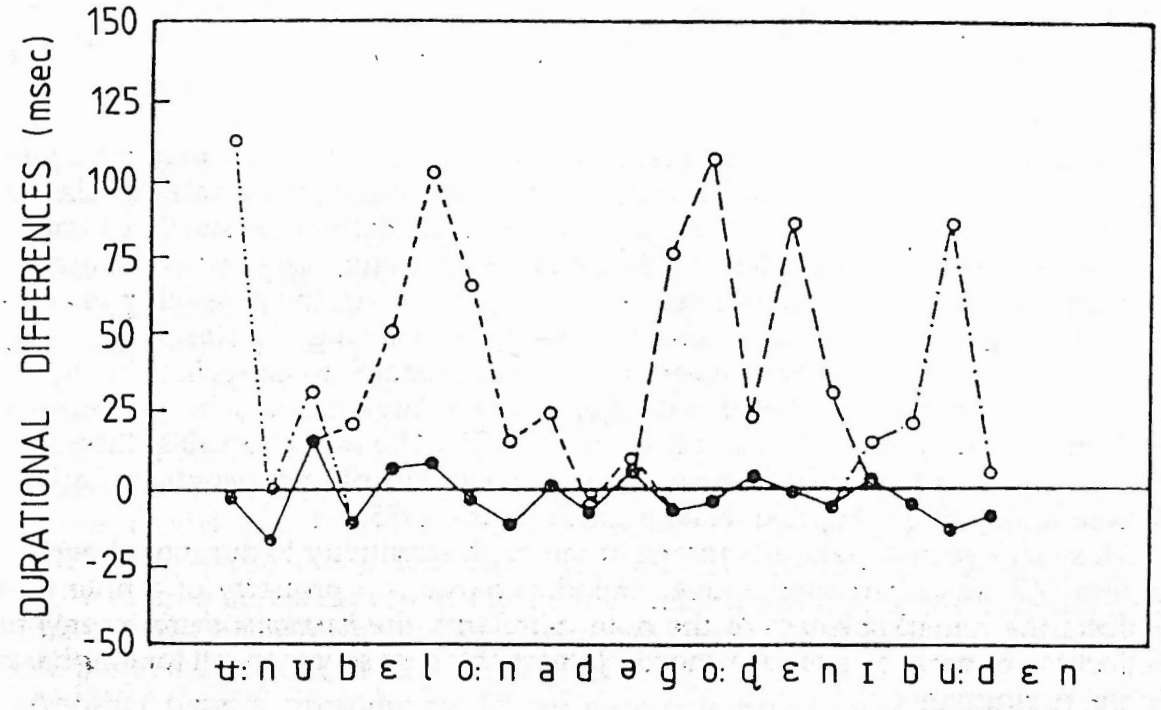

Figure 4 Durational differences between segments in emphatic and neutral utterances. The solid curve indicates the difference between segments in nonemphasized words of emphatic utterances and neutral production. The other curves pertain to the increase in duration of the words when pronounced emphatically.

performance (Friberg, Sundberg & Frydén, 1987). This marking seems to be an essential requirement in a music performance (Thompson *et al*, 1986). Measurements on music performance support the same assumption and also indicate that constituent marking takes place at different levels in the hierarchy (Todd, 1985).

Phrase marking is an instance of the general principle of marking constituents in a structure. This principle is often referred to as *grouping* which seems esssential in any type of communication. It seems that constituents at many different levels are marked. For instance, word boundaries are marked not only in speech, as mentioned above, but also in orthography with a space. Also in understanding the speech of a foreign language, a major step is to be able to detect the boundaries between words.

The fact that constituent marking appears not only in speech and music performance, but also in the orthography and the music score suggests that the marking of constituents is a paramount demand in many kinds of inter-human communication.

Choice of Acoustic Code

As both speech and music performance use acoustic signals for communication, it is interesting to compare the codes used. If speech and music performance use similar codes, the understanding of music requires a competence which is partly the same as that required for understanding speech. Thus, comparing the codes will shed some light on the basic requirements for understanding music.

Emphasis

As mentioned earlier, the main correlates for emphasis in speech are relatively greater pitch changes, increased durations, and, to some extent, greater vocal effort, as was illustrated in Figures 3 and 4. By and large, as pitch and duration in music are decided upon by the composer, much less leeway is left to the performer than to a speaker. However, while our sensitivity to differences in the duration of notes presented in isolation is modest, the sensitivity to minute perturbations of the duration appearing in a regular sequence of notes of similar duration is quite high. Under these conditions, perturbations as small as 10 msec can be detected (van Noorden, 1975). Thus, by arranging sequences of notes of similar duration, the composer seems to offer the player the possibility to communicate emphasis in terms of lengthening and shortening of notes.

Similar observations have been made with regard to speech. The lower boundary for perceiving durational differences has been found to be on the order of 10 msec (Huggins, 1972; Klatt & Cooper, 1975). The just noticeable difference has been shown to vary with the type of sound and its phonetic context (Carlson & Granström 1975b, Fujisaki, Nakamura & Imoto, 1975).

Musicians seem to take advantage of this high sensitivity to durational perturbation. As was mentioned above, melodic charge is a property of a note that reflects the remarkableness of the note. Similarly, the *harmonic charge* seems to reflect the remarkableness of a chord. Remarkableness seems to call for emphasis in the performance.

According to the performance rules, a high melodic charge is marked by increases in sound level, vibrato extent, and duration. The method is straightforward; the increment of sound level, vibrato extent, and duration is calculated

as a constant multiplied by the melodic charge of the individual note. The result is that notes with a high melodic charge sound emphasized. Similarly, increases in harmonic charge generate crescendos, and the associated increments in sound level are used for calculating the increases in duration and vibrato extent. According to formal listening experiments with musically trained subjects, the musical quality of a performance is raised if melodic charge is marked in this way (Thompson *et al*, 1986).

These examples seem to indicate that emphasis is signaled by adding duration, resulting in a slowing of the tempo. The same means are also used in speech. This slowing down of the tempo seems perceptually adequate; the listener is given more time to process the unexpected information.

Melodic charge and increases in harmonic charge are also reflected in the vibrato extent, as mentioned. Here, the parallel with speech is less obvious, but the following speculation is tempting. The perceptual system seems very sensitive to changes, e.g. in pitch, and one emphasis marker in speech is pitch change. Vibrato actually increases the rate of change of fundamental frequency, though without changing the mean perceived pitch. From this perspective, vibrato could be seen as an elegant way of exploiting pitch change for expressive purposes without changing the melodic patterns.

Constituent marking

In speech as in music performance, it is necessary to mark structural constituents at different levels. In speech, the most apparent example is the phrase and clause ending which is signaled by a lengthening of the last syllable or syllables. This way of announcing the ending of a constituent is common to most languages (Lindblom, 1979). Actually, final lengthening is so important for both Swedish and English that speech synthesized without such a rule is perceived as accelerating at the end of each clause.

In speech, constituents of many sizes, from paragraphs to words, are marked, not only with duration but also with other parameters like intonation, and vocal source settings. Also, micro-pauses are introduced at major syntatic breaks even if there is no need for breathing, e.g. after words followed by a period, a comma, or a semicolon. Durational data on these effects have been reported for several languages and are also formulated into coherent rule systems.

Prosodic models have an obvious importance in the general description of languages and find application in text-to-speech systems. Several of these models have been tested by Carlson, Granström & Klatt (1979), and the more elaborate models give advantages both in naturalness and intelligibility.

One durational description of Swedish is historically based on a tree structure of a sentence with phrase boundaries and syllables on separate branches. Support for this model was found in reiterant speech (Carlson & Granström, 1973; Lindblom & Rapp, 1973). The model had a cyclic rule that increased the final lengthening to an extent reflecting the hierarchical position of the boundary. A similar model has been found extremely productive in describing timing data from piano music performance (Todd, 1985).

Another type of prosodic model for speech is based on a general structure proposed by Klatt (1979). The rules have as input the inherent duration, which is the typical duration of the phoneme in a word-initial position before a stressed vowel. The second parameter is the minimal duration, which is a measure of the

phoneme's compressibility. Finally, a correction factor is used to calculate the duration. This factor is set depending on local and global parameters. This model has proven to be good at describing duration effects in running speech. The experiences from the music performance program reveal that a similar model, including restrictions on compressibility, would also be productive in music performance.

There seems to be a contradiction between these two models for describing duration phenomena. The first model is probably suited for a well-prepared reading of text with a high amount of speech pre-planning, while the other is typical of less planned speech with rules of a more local nature.

According to Todd (1985), the ending of a phrase is often played with a small retard while the beginning is played with a small accelerando. It is well known that the last notes of a piece are often played with a final retard (Sundberg & Verrillo, 1980). In the music performance program, the last note of a phrase is lengthened by 40 msec and terminated by a micro-pause. A sub-phrase termination is marked by a micro-pause only. In addition, there are a number of other rules that shorten and lengthen notes depending on the context, and these rules sometimes seem to serve the purpose of constituent marking. For instance, in combination with the marker of phrase endings, they actually sometimes generate small retards at phrase endings.

Why is the code for marking structural constituents similar in speech and music performance? A tempting hypothesis is that the code in music is imported from speech in this regard; as all music listeners have acquired a competence in decoding speech, it would be safe to use the same code in music performance. However, some languages, e.g. Danish, do not use final lengthening, and yet, musicians from these countries are obviously quite as competent musicians as their colleagues from other countries. This shows that the code used in music performance may not be borrowed from speech, but might lean on other kinds of common experience.

As far as the final retard is concerned, there is a striking similarity with the decreasing rate of footsteps in a stopping runner who keeps the step length and the braking force constant throughout the stopping process (Kronman & Sundberg, 1987). Under these conditions, the slowing down of the footsteps follows the same curve as the average retard in motor music from the baroque era. Thus, the final lengthening seems to allude to a well-known experience, namely that of stopping locomotion. We may speculate that the final lengthening in phrase endings are also faint allusions to locomotion. If so, the code would be very robust in the sense that anybody acquainted with locomotion is likely to know the code.

Outlook

We can discern three apparently very basic principles used in speech and music performance. The first one is increasing the difference between categories, which would facilitate communication by helping the listener to make correct identifications. A second principle is the *emphasis* which is called for by varying *predictability*. In speech, predictability would serve the purpose of making the message

robust. In music performance, on the other hand, this may or may not be the purpose; while speech is often required to function in a noisy environment, Western music is likely to be performed in less disturbed situations. In any event, it seems likely that it is the cognitive system that asks for varying degrees of emphasis. Perhaps, this system cannot digest long series of equally unexpected elements in communication. Emphasis is signaled acoustically in similar ways in speech and music.

Another common basic principle in speech and music performance is *constituent marking*. The parts that constitute blocks in the structure are marked in the acoustic realization, e.g. phrases and clauses in speech, and phrases and subphrases in music. This appears to reflect a requirement of the cognitive system, and is often referred to as the principle of *grouping*. Also, the acoustic code by means of which constituent marking is communicated is simpler in speech and music.

Why all these numerous parallels? What do they imply? The parallels are not astonishing. Both speech and music are examples of formalized inter-human communication by means of acoustic signals. Both must be devised for the same perceptual and cognitive systems. The limitations and capabilities of these systems must contribute importantly to the development of both speech and music.

Acknowledgments

The speech part of the work reported in this paper was supported by The Swedish Board for Technical Development (STU) Contract No. 84–3667 and the music performance part by the Bank of Sweden Tercentenary Foundation, Contract 84/171.

Notes

1. The authors' names are arranged in alphabetical order.

References

Bernstein, L. (1976) *The Unanswered Question*. Cambridge, Mass.: MIT Press.

Carlson, R., Erikson, Y., Granström, B., Lindblom, B & Rapp, K. (1975) Neutral and emphatic stress patterns in Swedish. In *Speech Communication*, Vol 2, G. Fant (ed.), 209–218, Stockholm: Almqvist Wiksell.

Carlson, R. & Granström, B. (1973) Word accent, emphatic stress, and syntax in a synthesis by rule scheme for Swedish. *STL–QPSR*, **2–3/1973**, 31–36.

Carlson, R. & Granström, B. (1975a) A phonetically-oriented programming language for rule description of speech. In *Speech Communication*, Vol 2, G. Fant (ed.), 245–253, Stockholm: Almqvist & Wiksell.

Carlson, R. & Granström, B. (1975b) Perception of segment duration. In *Structure and Process in Speech Perception*, A. Cohen & S. Nooteboom (eds.), 90–196, Heidelberg: Springer Verlag.

Carlson, R. & Granström, B. (1986) Linguistic processing in the KTH multi-lingual text-to-speech system. *Conference Record, IEEE–ICASSP, Tokyo*, 2403–2406.

Carlson, R. Granström, B. & Hunnicutt, S. (1982) A multi-language text-to-speech module. *Conference Record, IEEE–ICASSP, Paris*, 1604–1607.

Carlson, R., Granström, B. & Klatt, D. (1979) Some notes on the perception of temporal patterns in speech. In *Frontiers in Speech Communication Research*, B. Lindblom & S. Öhman (eds.), 233–244, New York: Academic Press.

Coker, C.H., Umeda, N. & Browman, C.P. (1973) Automatic synthesis from text. *IEEE Transactions in Audio Electroacoustics, AU–21*, 293–297.

Friberg, A., Sundberg, J. & Frydén, L. (1987) How to terminate a phrase: An analysis-by-synthesis experiment on a perceptual aspect of music performance. In *Action and Perception of Rhythm and Music*, A. Gabrielsson (ed.) 49–55, Publ. no. 55, Stockholm: Royal Swedish Academy of Music.

Fujisaki, H., Nakamura, K. & Imoto, T. (1975) Auditory perception of duration of speech and non-speech stimuli. In *Auditory Analysis and Perception of Speech*, G. Fant & M. Tatham (eds.) 197–200, New York: Academic Press.

Huggins, A.W.F. (1972) Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America*, **51**, 1270–1278.

Hunnicutt, S. (1985) Intelligibility versus redundancy – conditions of dependency. *Language and Speech*, **28**, 47–56.

Hunnicutt, S. (1987a) Acoustic correlates of redundancy and intelligibility. *STL–QPSR*, **2–3**, 7–14.

Hunnicutt, S. (1987b) Input and output alternatives in word production, STL–QPSR 2–3, 15–29.

Klatt, D.K. (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, *Journal of the Acoustical Society of America*, **59**, 1208–1221.

Klatt, D.K. (1979) Synthesis by rule of segmental durations in English sentences. In *Frontiers in Speech Communication Research*, B. Lindblom & S. Öhman (eds), 287–299, New York: Academic Press.

Klatt, D.K. & Cooper, W.E. (1975) Perception of segment duration in sentence contexts. In *Structure and Process in Speech Perception*, A. Cohen & S.Nooteboom (eds), 69–86,Heidelberg: Springer Verlag.

Knopoff, L. & Hutchinson, W. (1983) Entropy as a measure of style: The influence of sample length, *Journal of Music Theory*, **27**, 75–97.

Kohler, K.J. (1981) (ed.) "Temporal Aspects of Speech Production and Perception", *Phonetica*, **38**, 1–3.

Kohler, K.J. (1986) (ed.) "Prosodic Cues for Segments", *Phonetica*, **43**, 1–3.

Kronman, U. & Sundberg, J. (1987) Is the musical retard an allusion to physical motion? In *Action and Perception of Rhythm and Music*, A. Gabrielsson (ed.), 57–68, Publ. no. 55, Stockholm: Royal Swedish Academy of Music.

Krumhansl, C.L. & Kessler, E.J. (1982) Tracing the dynamic changes in perceived tonal organization in spatial representation of musical keys. *Psychological Review*, **89**, 334–368.

Lehiste, I. (1975) The phonetic structure of paragraphs. In *Structure and Process in Speech Perception*, A. Cohen & S.Nooteboom (eds), 195–203, Heidelberg: Springer Ferlag.

Lerdahl, F. & Jackendoff, R. (1983) *A Generative Theory of Tonal Music*. Cambridge, Mass: MIT Press.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172–187.

Lindblom, B. (1979) Final lengthening in speech and music. In *Nordic Prosody*, E. Gårding, G. Bruce & R. Bannert (eds), 85–101, Lund: Travaux de l'Institut de Linguistique de Lund XIII.

Lindblom, B. & Rapp, K. (1973) Some temporal regularities of spoken Swedish. University of Stockholm, Dept. of Linguistics, Publ. No. 21.

Lindblom, B. & Sundberg, J. (1970) Towards a generative theory of melody. *Svensk Tidskrift för Musikforskning* (Swedish Journal of Musicology), **52**, 171–181.

van Noorden, L.P.A.S. (1975) *Temporal Coherence in the Perception of Tone Sequences*. Dissertation, Technical University, Eindhoven.

Sloboda, J.A. (1977) Phrase units as determinants of visual processing in music reading. *British Journal of Psychology*, **68**, 117–124.

Sundberg, J. (1978) Synthesis of singing. *Svensk Tidskrift för Musikforskning*, (Swedish Journal of Musicology), **60**, 107–112.

Sundberg, J. (in press), Synthesis of singing using a computer-controlled formant synthesizer, manuscript to be published by the Music Department, Stanford University

Sundberg, J. & Lindblom, B. (1976) Generative theories in language and music descriptions. *Cognition*, **4**, 99–122.

Sundberg, J. & Verrillo, V. (1980) On the anatomy of the retard: A study of timing in music. *Journal of the Acoustical Society of America*, **68**, 772–779.

Sundberg, J., Askenfelt, A. & Fryden, L. (1983) Musical performance: A synthesis-by-rule approach. *Computer Music Journal*, **7**, 37–43.

Thompson, W.F., Friberg, A., Frydén, L. & Sundberg, J. (1986) Evaluating rules for the synthetic performance of melodies. *STL–QPSR*, **2–3**, 27–44.

Todd, N. (1985) A model of expressive timing in tonal music, *Music Perception*, **3**, 33–57.

Winograd, T. (1968) Linguistics and the computer analysis of tonal harmony. *Journal of Music Theory*, **12**, 2–49.