SEPARATE PITCH AND RHYTHMIC PATTERNS IN SYNTHETIC SPEECH AND MUSIC

JAN TRO

The Norwegian Institute of Technology, Acoustics
N-7034 Trondheim-NTN, NORWAY

ABSTRACT

For Text-to-Speech synthesizers normally the text is manually prepared to add some prosodic information. This is rather time consuming and requires an expert user of the system. Even if speech and music seem to be closely related as acoustical signals, we need different procedures to add melodic and prosodic information into a music and a speech synthesizer, respectively. The strong relation of syllable duration, pitch and sound level in Norwegian dialects seem to complicate the procedure of adding humanlike prosody to synthetic speech.

This paper deals with aspects of similarities between speech and music signal. The aim is to contribute to the simplification of adding Norwegian prosodic features to speech synthesizers.

INTRODUCTION

Prosodic parameters as syllable duration, pitch and sound level are closely linked variables in Norwegian dialects. This fact complicates the procedure of adding humanlike prosody to synthetic speech.

As long as the intonation in some Norwegian dialects is very important to establish the correct meaning of an utterance, we need simple methods for adding sufficient prosodic information to synthesizers.

If we consider both music and speech as elements in the process of human communication, both signals may be described by a set of acoustical parameters in the frequency and the time domain. It is very easy to discuss musical details (pitch, rhythm, level) as near independent parameters.

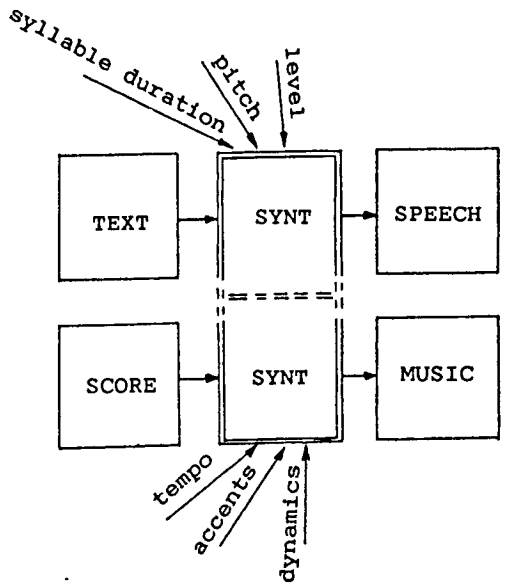

Figure 1: Sound synthesis with additional information.

Even if speech and music seem to be closely related as acoustical signals, we need different procedures to add melodic and prosodic information into a music and a speech synthesizer, respectively.

Could it still be possible to add prosodic information factor by factor, - in order to obtain increased intelligibility, understanding and naturalness?

SPEECH AND MUSIC

It is not a quite new idea to compare features in speech and music. In 1916 Alnæs [ref. 1] used the standard musical notation in his book to described the variation of pitch, - with minor success.

Nevertheless there is obviously a need for some kind of graphic notations in order to describe prosodic information in detail. The more common used system with dots and lines for unstressed and stressed syllables seems to give a suitable precision in the description of pitch variation. This system is still very close to the musical notation with the traditional five lined staff with dots (tones) of different duration.

Rakowski [ref. 3] has considered music as one part in the process of human communication. In this way it is possible to discuss music on the basis of information theory. The distinction of sensory features which differentiate musical intervals seems to be bases on the same principle as with vowels of speech. That means we are dealing with a set of cognitive prototypes of tone color in the case of vowels and prototypes of interval size in the case of musical intervals. The reason why the number of phonological units of speech and music is not very high (30-40 phonemes, 12 equally tempered intervals per octave), is probably more because of the limited capacity of memory than due to the discriminatory properties of hearing.
This approach is indeed very promising when comparing speech and music signal.

However, it is possible to make a shorter way to this comparison saying that the intonation of speech, the "melody of speech", is the aspect which comes closest to music [ref. 7]. This is an acceptable assumption as long as "music" is restricted to comprise melodic lines only. Let us have a closer look at basic parameters in melodic lines.

Pitch
The limited collection of notes in traditional western melodic lines consist of a set of fixed tones positioned on the well-known five lined staff. It is very tempting to use a similar notation system for the pitch variation of syllables. One model is proposed by Sirnes [ref. 4]. However, some problems will arise due to the fact that music notation is normally designed for instruments with fixed pitch. All your practicing and rehearsals will lead to an even closer connection to the accepted collection of tones. This is a genuine situation for the performance of classical western music. The attempt to use a set of fixed tones in speech will give the impression of trying to recite in a singsong manner.

Sirnes has proposed some modifications for his model in order to implement nonlinear effects (nonsymmetrical syllable transitions, avoid stationary pitch, etc.). This will probably contribute to increased naturalness of the synthesis.

As far as music concerns, removing the fixed pitch is fatal to the perception of melody. Even the fixed intervals is a basic property of music in contradiction to normal speech.

Is it possible to deal with pitch as an independent parameter?

It is not possible in music because the combination of time and pitch information is necessary for the correct perception of melodic lines. Although the intelligibility of speech will remain unaffected, the lack of time information will dramatically decrease the naturalness of Norwegian dialects. Analysis of Norwegian native dialects show strong bindings of syllable pitch and duration as prosodic elements. We have to remember that pitch changes may even change the meaning of the utterance.

Duration
Deviations in the time domain will affect music and speech differently. In melodic lines both relative time and absolute duration of notes are very important in order to obtain a proper rhythmic percept. Any changes in the microstructure of the notes (short/long attack time, decay time, sustain time and release time) will normally affect the perception of timbre keeping the melody still intact.

Deviations of syllable start time and duration will not have a corresponding fatal effect. As minor deviations will lead to decreased naturalness, only larger deviations may result in decreasing intelligibility.

Is it possible to deal with tone/syllable duration as an independent parameter?

The strong bindings of pitch and duration in the constructions of melodic lines is well-known and do not invite us to get rid of the pitch information. However, if we need to analyse or synthesize the rhythmic patterns it is a good idea to concentrate on the time domain exclusively. This holds for speech synthesis as well.

If we remove the pitch information (meaning the voiced sounds) the utterance may still be accepted as natural, namely as whisper! In this way it is possible to deal with syllable duration in forming rhythmic patterns of utterances without being confused by the pitch contour. This principle has been tested and reported by Ottesen [ref. 2].

Intensity
The recognition of melodic lines are nearly affected at all by variation in the tone intensity. However, as part of the dynamic musical structure it is very important.

In Norwegian dialects syllable intensity plays a minor role compared to pitch and time information in the prosodic structure.

Some valuable information on the construction of the syllable envelope curves and the transition from one curve to another, may be found in the area of musical synthesis.

Speech quality
The discussion and definition of speech quality is of vital importance for the evaluation of synthetic speech. This discussion has to include aspects of intelligibility, sense of utterance, and naturalness. We have to define what is an acceptable, expected, and a sufficient level of quality.

As we have seen, the quality of spoken Norwegian dialects is affected by variations in pitch, duration and intensity. Even feelings may be expressed by some changes in the prosodic features. In some cases large prosodic differences (like in different Norwegian dialects) may be evaluated as natural. But only minor prosidic variations may cause in a fatal decrease in naturalness when it is spoken by a non-native tongue. This means that adaptational processes will influence the evaluation of quality.

It is important to establish references for all the different levels of synthesized speech quality.

LISTENING TESTS

Experiences of musical listening tests on pitch and rhythm has been an advantage in designing our laboratory tests.
Through different experiments we try to isolate independent prosodic elements which can easily be added separately to the speech synthesizer in order to obtain more humanlike speech sounds.

Listening tests has included recognition of melodic lines, separation of pitch and time information in melodic lines, effects of prosodic variations in a Norwegian dialect, and pitch changes as a result of intensity variation.
Results from these tests will be presented.

REFERENCES

[1] Alnæs, I.:
Norsk sætningsmelodi,
(Norwegian Prosody)
Kristiania, 1916. (in Norwegian)

[2] Ottesen, G.:
Adding natural prosody to a phoneme synthesiser,
11th PCPhS, Tallin, 1987.

[3] Rakowski, A.:
Acoustics of music and acoustics of speech. Some common factors.
Proceedings of 23rd Acoustical Conference on Physiological and Psychological Acoustics, Acoustics of Speech and Music. Ceske Budejovice, 1984.

[4] Sirnes, G.:
Analyse og syntese av norsk setningsprosodi. (Analysis and synthesis of Norwegian Prosody) Thesis.
The Norwegian Institute of Technology
Trondheim, 1986. (in Norwegian)

[5] Tro, J.:
Aspekter ved lytting til levende og teknisk gjengitt musikk. (aspects of Listening to Live and Recorded Music)
ELAB memo 44-AN85030.
Trondheim, 1985. (in Norwegian)

[6] Vanvik, A.:
Norsk fonetikk. (Norwegian Phonetics)
The University of Oslo,
Oslo, 1979, (in Norwegian)

[7] Vanvik, A.:
Reflections on the relation between speech and music.
In Fretheim (ed): Papers from a Symposium. Nordic Prosody II.
Trondheim, 1980.