DATA BANK ANALYSIS AND NASAL SYNTHESIS

Lennart Nord

Department of Speech Communication and Music Acoustics,
Royal Institute of Technology (KTH), Box 70014,
S-100 44 Stockholm, Sweden.

1. INTRODUCTION

This is a brief summary of recent work on segmental analysis and synthesis. Specifically, the aim is to improve the quality of nasal sounds in our text-to-speech synthesis system, developed at the Speech Department in Stockholm [1].
Our acoustic-phonetic work is presently making use of some new software facilities, developed at the Department. Another recent feature is the support from our speech data bank that is being structured. A number of earlier reports have dealt with different aspects of the data bank project [2-6].
For expanded versions of this report, see [7,8].
Theoretical and modelling work has given us a good insight into the acoustics of nasals and the effects of nasal coupling [9-14]]. Still, perceptual evaluations of synthesis systems, usually give high error rates for nasals compared to other speech sounds.

2. NASAL SYNTHESIS

2.1 Serial synthesizer

Synthesis work at the Speech Communication Department traditionally exploits a serial type of terminal analog formant synthesizer, the OVE III as a part of our text-to-speech system [15]. There is thus a fixed relation between formant frequencies and relative spectral levels, which has to be taken into account in the spectral matchings. As there are numerous extra resonances and zeroes to be found in nasal spectra, the matchings are difficult.

2.2 Evaluation of present text-to-speech system

Systematic listening tests have been performed during the years, using /VCV/ words with V-V = /a-a, I-I, U-U/. As a complement a separate test was made with /CV:Ca/ words with /V:/ = /a:, i:, u:/ and the consonant C = all combinations of /m,n,l/, giving words like *mila, lama, nola.* An analysis of the nasal errors show errors in manner as well as in place of articulation. Specifying the probable sources of error will give a representative repertoire of the nasal acoustic cues. In Table I some of the typical errors are listed, together with a brief comment of source of error. Error scores are not given here as they fluctuate considerably with sound quality and listener sophistication.

TABLE I

	Probable Main Source of Error
/ni/ -> /mi/	formant transition
/no/ -> /mo/	formant transition and poor nasal resonance
/mo/ -> /vo/	transition rate and poor nasal resonance
/ŋi/ ->/ji, gi/.	intensity level and poor nasal resonance

2.3 Continuing synthesis work

2.3.1 Software facility.
We are presently using a convenient software tool, RULSYS, developed at the Speech Department. The program package will allow us to select a natural speech sample,

make an FFT analysis of it and display a spectrogram on the monitor together with the rule synthesis output in the form of synthesis parameter tracks. The resulting synthesis can be played back and compared to the sound of the natural speech sample. At any point along the time axis a spectral section can be derived both of the synthesis output and the speech signal, a necessary component of the matching procedure.

2.3.2 Pattern matching and testing synthesis strategies.
The software facility is used in two ways. Firstly, to generate a set of detailed pattern matchings of vowel-nasal-vowel syllable, trying to achieve synthetic copies with as good quality as possible. Once we get perceptually and acoustically acceptable matches, we start degrading the parameter tracks to establish the relative importance of the acoustic cues. Formant transitions, transition rates, bandwidth values, nasal segments as opposed to nasalized vowels, etc are evaluated. The other line considers the testing of various synthesis strategies. As there are some built-in restrictions of the synthesizer, this will give us a quick feedback regarding some of the principles that should be used in synthesizing nasals. The nasal segment can thus consist of the single nasal resonant, the vowel formants with proper bandwidth modifications etc. For earlier work on nasal synthesis, see [16,17]

3. A SWEDISH SPEECH DATA BANK

3.1 Purpose

An important component in improving the text-to-speech system is the supply of real speech data and especially the possibility to analyse large materials. Our data bank is labelled at a phonemic level, and a systematic search can be made through the material by the use of phoneme strings. Our long-term goal is to include ten males, ten females and ten children, but so far only four males and one female have been recorded reading the entire material, and only two of the male readings have been partially labelled. Eventually, it will be very interesting to get data on suffixes from running text material, that frequently consists of vowel-nasal syllables.

3.2 Recording and labelling

The speech is digitized at a sampling frequency of 16 kHz and edited on a computer by means of a speech wave editing program. After editing, the material is stored on 80 Mbytes computer discs. First, a phoneme string is generated from the text-to-speech system, Carlson & Granström [1], using the orthographic strings as input. The phoneme strings, including prosodic markers, are synchronized to the speech wave by means of the speech editing program. To facilitate the placing of labels, part of an automatic speech recognition system, Blomberg & Elenius [18] can be used. By using computer generated spectrograms as a complement, the automatically produced segmentation is adjusted by moving the marks along the oscillographic tracing on the computer screen, while listening to the signal. This process is very time-consuming, and as we at the same time are trying to improve our segmentation criteria there is no way to make this phase of the work automatic. When finally an accepted labelling of the material is attained, the label files will be stored together with the speech files on the 80 Mbyte discs and the material is ready for usage.

3.3 Analysis

The analysis will also rely on the text-to-speech program that generated the phoneme strings, in an automatic search procedure. The search is made by specifying the phoneme or the class of phonemes that are of interest, together with prosodic markers. The result can either consist of automatically derived measures of segment durations, spectral envelopes or simply a subset of phones for visual inspection and subsequent later analysis. Depending on the detailed prosodic marking, contextual specifications may include details of stress levels and position within the word, thus, for example enabling an analysis of unstressed verses stressed syllables. As an example of data bank usage this routine has been performed on another set of speech material by Carlson & Granström, who reported on duration measures of Swedish and American vowels and consonants, all of which were derived in this fashion [4].

3.4 The phonemic principle

Our label string is thus basically phonemic, and allophonic variations are not taken into account, unless they are clearly the result of some higher level rule (e g the dialectal pronunciation of proper names). This means that any allophonic variation will be seen in the distribution of analysis data. If we for example ask for a formant plot of all stressed /ö/ phonemes, we will probably get a bimodal distribution in the formant plane, in this particular case due to the fact that there are two contextually dependent /ö/ allophones in Swedish. There are two reasons for this 'phonemic' approach. Firstly, it will lessen the demands on part of the labeller, who otherwise could spend an infinite amount of time, trying to chose among IPA symbols and diacritics. Secondly, as we are trying to improve the quality of synthetic speech, we are usually in a situation where, starting from a phoneme string, the goal is to generate a satisfactory acoustic output. The statistical distribution of how certain phonemes strings will be modified by coarticulatory processes and realized is easily found by making a search in the material, using the string of interest.

3.5 Texts

Our choice of texts tries to satisfy a number of demands from a phonetic-acoustic point of view. The material consists of speech sounds in various contexts, ranging from isolated phonemes to text passages. Vowels, consonants, syllables and words pronounced in isolation with a lexical stress are included as a means to establish a basic acoustic mapping of the speakers, and also enabling studies of reduction properties when going from careful pronunciation to more relaxed. Different types of text: isolated sentences, newspaper articles of different complexity (politics, general news bulletin, speech research), and part of a novel by a Swedish novelist, complemented by one page that is read with some variation in style (normal, clear, weak, strong) and a short piece of dialogue from the same novel constitute the entire text material.

3.6 Choice of speakers

Speakers are chosen according to two principles. On the one hand we are interested in trained speakers, such as radio announcers, that can read well and come close to a natural sounding, almost spontaneous speaking style. On the other hand we are also interested in normal, non-professional speakers, and especially for recognition work it would be fatal not to include variation for the analysis. In the work on Swedish prosody by Fant, Nord and Kruckenberg [6] and Fant and Kruckenberg [19], one of the aims is to understand the personal variations in reading styles.

3.7 Principles of segmentation and labelling

Most of the labelling difficulties occur in the text passages. There is indeed a great difference between this kind of material and isolated sentence material. In the fluently read texts strong reductions often occur, and in contrast to the isolated sentences, large parts of the text, sometimes whole phrases lie out of focus.

5. REFERENCES

[1] R CARLSON & B GRANSTRÖM,"Linguistic Processing in the KTH multilingual text-to-speech system", Conference record, IEEE-ICASSP, Tokyo (1986)

[2] R CARLSON & B GRANSTRÖM,"Rule Controlled Data Base Search", STL-QPSR 4/1985, p 29 (1985)

[3] R CARLSON & B GRANSTRÖM, "Swedish Duration Rules Derived from a Sentence Data Base", STL-QPSR 2-3/1986, p 13 (1986)

[4] R CARLSON & B GRANSTRÖM, "A Search for Durational Rules in a Real-Speech Data Base", Phonetica 43, p 140 (1986)

[5] G FANT, L NORD & A KRUCKENBERG,"Individual Variations in Text Reading . A Data-Bank Pilot Study", STL-QPSR 4/1986, p 1 (1986)

[6] G FANT, L NORD & A KRUCKENBERG,"Segmental and Prosodic Variabilities in Connected Speech. An Applied Data-Bank Study", Proc XXIth ICPhS, Tallinn, USSR, Vol.6, p 102 (1987)

[7] L NORD, "Acoustic-Phonetic Studies in a Swedish Speech Data Bank", paper presented at the FASE symposium, Edinburgh (1988) and forthcoming STL-QPSR

[8] L NORD, "Synthesis of Nasal Sounds in a Text-to-Speech Framework", paper presented at the FASE symposium, Edinburgh (1988) and forthcoming STL-QPSR

[9] A HOUSE & K STEVENS, "Analog Studies of the Nasalization of Vowels", J Speech Hear Dis 21, p 218 (1956)

[10] S HATTORI, K YAMAMOTO & O FUJIMURA,"Nasalization of Vowels in Relation to Nasals", J Acoust Soc Am 30:4, p 267 (1958)

[11] O FUJIMURA,"Spectra of Nasalized Vowels", Q Prog Rep 58, Res Lab Electron M I T, p 214 (1960)

[12] O FUJIMURA & J LINDQVIST-GAUFFIN,"Sweep-Tone Measurements of Vocal-Tract Characteristics", J Acoust Soc Am 49:2, p 541 (1971)

[13] J LINDQVIST-GAUFFIN & J SUNDBERG,"Acoustic Properties of the Nasal Tract", Phonetica 33, p 161 (1976)

[14] S MAEDA,"The Role of the Sinus Cavities in the Production of Nasal Vowels", Proc of ICASSP 1982, Paris, France, p 911 (1982)

[15] R CARLSON, B GRANSTRÖM & S HUNNICUTT,"A Multi-Language Text-To-Speech Module", Proceedings IEEE-ICASSP, Paris (1982)

[16] L NORD,"Perceptual Experiments with Nasals", STL-QPSR 2-3/1976, p 5 (1976)

[17] L NORD,"Experiments with Nasal Synthesis", STL-QPSR 2-3/1976, p 14 (1976)

[18] M BLOMBERG & K ELENIUS,"Automatic Time Alignment of Speech with a Phonetic Transcription", STL-QPSR 1/1985, p 37 (1985)

[19] G FANT & A KRUCKENBERG,"Some Durational Correlates of Swedish Prosody", paper presented at the FASE symposium, Edinburgh (1988)