Segment level improvements in a preliminary female text-to-speech system

Inger Karlsson and Lennart Neovius
Dept of Speech Communication and Music Acoustics, KTH,
Box 70014, S-100 44 Stockholm, Sweden

Some experiments to improve the naturalness of synthetic female speech will be presented. Synthesis improvements have been made mainly on the segmental level by including voice source variations and more complex articulation modelling. Source, formant and pole/zero parameter data generated by inverse filtering have been interpreted and incorporated into the female text-to-speech rule system. In particular, the quality of consonants with complex articulations and vowel-consonant transitions is enhanced. A diagnostic intelligibility test was performed on the preliminary rule system. Results will be discussed in terms of error sources and possible solutions. On-going work includes fine-tuning of LF-source and pole/zero parameter control, based on these results.

INTRODUCTION

Text-to-speech synthesis systems have, during the last decade, incorporated more sophisticated speech production models, including more elaborate glottal source representations (Fant, Liljencrants and Lin 1985, Klatt and Klatt 1991). The inclusion of voice source variations in the speech synthesis has, in earlier experiments, been shown in informal listening tests to improve the quality (Carlson, Granström and Karlsson 1991). An implementation of parameter-controlled zeroes in the vocal branch gives, according to our experiments, a further improvement of the naturalness of the synthetic speech (Karlsson and Neovius 1993). Still we lack data and experience needed for controlling these new models. Research is now focusing on collecting data and on formalising these new data into rules. This is necessary especially for the production of female and child voices. More data about natural speech will also make it possible to produce different voice qualities and speaking styles.

Detailed analysis of natural female speech, using inverse filtering technique, has widened our knowledge of voice source behaviour and of vocal tract characteristics (Karlsson 1990, 1992). In inverse filtering, a detailed description of voice source behaviour is obtained by cancelling each formant with an inverse formant filter, with the corresponding mid-frequency and bandwidth. Analysis data, including voice source, formants and poles and zeroes, on VCV-sequences from female speech will be presented in another paper at this conference (Karlsson 1994). These data are formalised and used in the synthesis rules and have enhanced the quality of consonants and vowel-consonant transitions.

METHOD

RULSYS/HISYS (Carlson, Granström and Hunnicutt 1990) has been used to hand-edit rule-generated speech to incorporate the analysis data. The data are then generalised as segmental rules and implemented in the female text-to-speech system. Parameters are used to control the GLOVE speech synthesiser.

We have used the male text-to-speech rule system as a starting point, giving us the global structure and framework. The male system was originally based on the old OVE synthesis configuration (Nord 1976, Carlson and Nord 1991). A female text-to-speech system can not be attained as a direct transformation, though. All basic definitions need to be re-written. During the development of the female speech synthesis, new features have been added to the GLOVE synthesiser. The inclusion of the LF-model and a pole /zero pair, necessitates a whole new set of rules, and also entails the re-writing of some old rules.

The work has been focused on getting a complete inventory of consonants for the female text-to-speech system. Modifications indicated by an earlier test were included in the rules. In this paper, we will not discuss the rules in detail, but provide /l/ as one example of presently implemented parameter controls. Further examples are given in an earlier paper, Karlsson and Neovius (1994).

For /l/, two different aspects were formalised. A pole/zero pair was specified for the whole segment. The specifications were the same for /i/ and /a/ contexts, for /u/-context the pole/zero pair as well as the formant values were lower. The voice source contained less high frequency energy in /l/ than in the surrounding vowels. The voice pulse was set to be more symmetrical than in the vowels, which added extra energy in the lowest harmonics. The excitation energy was reduced for the first 20 ms of the /l/ which improved the naturalness notably. A representation of a natural and an old and a new synthetic production of /ili/ are shown in Figure 1.

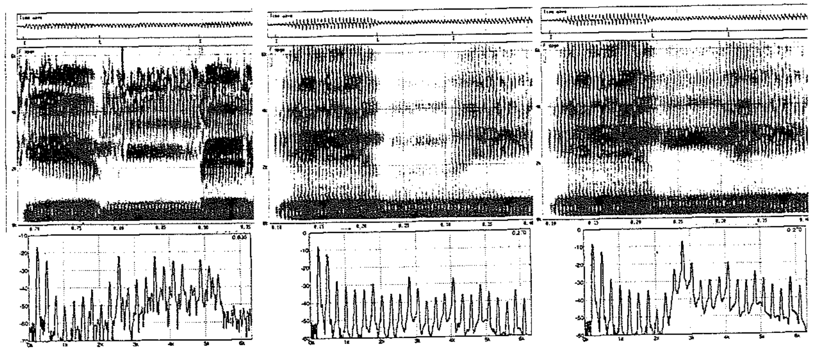

Figure 1. *Spectrograms and spectral sections of /ili/ produced by a female speaker and by old and new synthesis rules. The old synthesis does not contain the pole/zero pair, all other parameters are the same for the old and the new synthesis. The spectral sections are taken from the middle of the /l/.*

LISTENING TEST

A diagnostic test was performed on the preliminary rule system. The error rates achieved in this test were high, which was anticipated. The test was merely intended to give guidelines for future developments.

The participants were asked to correctly identify the consonant in a VCV stimuli, following the SAM procedure (Goldstein and Till 1992) but also including the retroflex allophones to get data on the complete Swedish consonant inventory. This resulted in 23 consonant phonemes. Each consonant was presented in three different short vowel contexts, /a_a/, /i_i/, /u_u/. The stimuli were produced with Swedish accent 2 and they were presented once in a forced-choice test. The vowel context was indicated in the answering form. 10 subjects participated in the test. Overall error rate was 48% for correct identification. When only manner of articulation was considered, the error rate was 20%. When the retroflexes were excluded, overall error rate decreased to 31% for correct identification. The corresponding overall error rate for natural male speech is 5.6%, and for the latest revision of the KTH male speech synthesis 8.7%, (Goldstein and Till, 1992). Measuring manner of articulation, the error rate was 18% on average with the retroflexes excluded.

Different types of error sources were identified in an examination of the consonant confusions. Some consonants gave a very high confusion rate, in particular /h/ and the retroflex consonants. One possible reason for the low intelligibility of the consonant /h/ is that we so far lack knowledge on how to control the noise adduction parameter by rule. The retroflex consonants were nearly always perceived incorrectly. This is mainly due to that they are comparatively rare in the language and that the voiced retroflexes hardly ever occur after a short vowel as in the test.

There was a trend to perceive the unvoiced stops as voiced, which was particularly strong for /a/-context. A possible explanation is that the voice off-set of the preceding vowel is too slow. When the speed parameter was modified, the intelligibility was improved. A further improvement was achieved by introducing adduction noise and a steeper tilt of the voice source spectrum in the later part of the preceding vowel. For labial stops, male values had survived. These were changed to correspond better with female data.

In velar consonants, the place cue was not strong enough. Moving the F2 and F3 locus closer to each other, and for stops also lowering the spectral point of gravity of the burst, gave better velars.

The nasals, that in an earlier test using hand-edited stimuli (Karlsson and Neovius 1994) were rated equally good in quality to the old nasals, appear to have lost their nasal cue. Confusions with non-nasals are particularly frequent for the /u/ context. One possible explanation is that the vowel-nasal transition contained noise, originating from the voice source. The control of the LF-model in the present implementation, depending on the fundamental frequency to sampling rate quotient, can be very demanding and errors are probably due to the fact that the precision needed has not been achieved so far.

CONCLUSIONS

In this report we have presented the progress of the KTH female text-to-speech synthesis. The work has been carried out within the text-to-speech development framework with the aim of generating more natural sounding synthetic speech. The experiments have focused on female speech as the quality can not be improved from the present status without the inclusion of a versatile voice source and more complex articulation possibilities.

Inverse filtering with matching against the LF-model supplies adequate analysis data for the formulation of synthesis rules. Analysis data on consonants and vowel-consonant transitions have been formalised and incorporated within the rule system. The inclusion of an extra pole/zero pair in the vocal branch gives us a possibility to generate more natural synthetic speech. Formalising the data to rules is still at an initial stage.

A diagnostic test has been performed to give guide-lines for further improvements. The test indicated some systematic errors. At the present stage, we have identified these errors and have corrected some of them.

Certain features tested in the system gave good results. The use of glottal source control to produce the voice bar in the voiced stops gives the desired responses. Very few confusions with other manners of articulation occurred. The production of the voice bar using glottal parameters makes it possible to generate a more or less reduced occlusion. The introduction of a pole/zero pair and an amplitude dip resulted in an improved /l/.

ACKNOWLEDGEMENTS

This work has been supported by FRN, NUTEK, HSFR, KTH and the European CEC-ESPRIT project SPEECH MAPS.

REFERENCES

Carlson R., Granström B., Hunnicutt S. (1990): "Multilingual text-to-speech development and applications", *Advances in speech, hearing and language processing* (edited by B Ainsworth), Vol. 1, pp. 269-296, JAI Press Ltd., London

Carlson R., Granström B., Karlsson, I. (1991): "Experiments with voice modelling in speech synthesis", *Speech Communication,* Vol. 10, pp. 481-490

Carlson, R. and Nord, L. (1991): "Positional variants of some Swedish sonorants in an analysis-synthesis scheme", *Journal of Phonetics*, Vol. 19, pp. 49-60.

Fant, G., Liljencrants, J., Lin, Q. (1985): "A four-parameter model of glottal flow.", *STL-QPSR* 4/1985, pp. 1-14.

Goldstein, M., Till, O., (1992): "Is % overall error rate a valid measure of speech synthesiser and natural speech performance at the segmental level?", Proc of ICSLP 92, pp. 1131-1134.

Karlsson, I., (1990): "Voice source dynamics for female speakers" Proceedings of the 1990 International Conference on Spoken Language Processing, Kobe, pp. 69-72.

Karlsson, I. (1992): "Consonants for female speech synthesis", *Proc. of 1992 Int. Conf. on Spoken Language Processing*, pp. 491-494

Karlsson, I. (1994): "Acoustic description of consonants in female speech; voiced consonants", this volume

Karlsson, I., and Neovius, L. (1993): "Speech synthesis experiments with the GLOVE synthesiser", Proc. of Eurospeech '93, pp. 925-928.

Karlsson, I., and Neovius, L. (1994): "VCV-sequences in a preliminary text-to-speech system for female speech ", *STL-QPSR* 1/1994 (forthcoming).

Klatt, D., and Klatt, L. (1990): "Analysis, synthesis and perception of voice quality variations among female and male talkers". *JASA,* Vol. 87, pp. 820-857

Nord, L. (1976): "Experiments with nasal synthesis", *STL-QPSR* 2-3/1976, pp. 14-19