Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

# Experiments with voice modelling in speech synthesis

Carlson, R. and Granström, B. and Karlsson, I.

journal: STL-QPSR
volume: 31
number: 2-3
year: 1990
pages: 053-061

http://www.speech.kth.se/qpsr


KTH Computer Science and Communication

# EXPERIMENTS WITH VOICE MODELLING IN SPEECH SYNTHESIS*

*Rolf Carlson, Björn Granström, and Inger Karlsson***

**Abstract**

Some experiments with voice modelling using recent developments of the KTH speech synthesis system will be presented. The new synthesizer, GLOVE, is an extended version of OVE III. It contains an improved glottal source built on the LF-model and some extra control parameters for the voiced and noise sources and an extra pole/zero-pair in the nasal branch. Furthermore, the present research versions of the KTH text-to-speech system include possibilities for interactive manipulations at the parameter level with on-screen reference to natural speech. The synthesis system constitutes a flexible environment for voice modelling experiments.

The new synthesis tools and models were used for synthesis-by-analysis experiments. A sentence uttered by a female speaker was analyzed and a stylized copy was made using both the old and the new synthesis system. It was found that with the new system, the synthetic copy sounded very similar to the natural utterance.

## INTRODUCTION

The need for voice variations is apparent in different speech synthesis applications, such as voice prosthesis and translating telephony. Speaking style variations are an important means of discriminating information of different kinds, Bladon et al. (1987). Data on speaker variability is now being accumulated. Fant, Kruckenberg, & Nord (1990) have investigated speaker variations in the context of a multi-talker speech data base. More detailed analysis of voice source dynamics have been studied by Gobl (1988) and by Karlsson (1988).

In this presentation we want to describe some recent experiments with voice modelling. We have used speech synthesis as a research vehicle to study global effects, voice transformations, and more individualized transforms implemented by changes in definitions and rules in the text-to-speech system. The present research versions of the KTH text-to-speech system and the possibility for interactive manipulations at the parameter level with on-screen reference to natural speech constitute a flexible environment for such experiments. Special effort is invested into the creation of a female voice. Transformations by global rules of male parameters are not judged to be sufficient. Changes in definitions and rules are made according to data from a natural female voice.

We have recently implemented a more realistic voice source, an expansion of the LF model. The spectral properties of this new voice source and the possibility of dynamic variation have proved to be essential for modelling a female voice.

Our approach has been to make a stylization of individual utterances spoken by speakers with different voice characteristics. In this process we have started with rule-generated parameters and adjusted the target values according to the different voices, using the possibility to overlay a spectrogram of natural speech on the speech synthesis parameter traces. Results from inverse filtering have been used in setting the appropriate voice source parameters. Typically one or two specifications per speech sound has been used for each

---

*This contribution was presented as an invited tutorial paper to the ESCA workshop on Speaker Characterization in Speech Technology, Edinburgh, June 26-28, 1990.
**Names in alphabetic order.

vocal tract and source parameter, i.e., in contrast to the approach taken by Pinto, Childers, & Lalwani (1989), we are not trying to model the human speaker frame-by-frame but rather to make a stylization that later on can be generalized when formulating rules for the different speakers.

In our paper, we will describe an extended synthesizer GLOVE, compared to the standard OVE III (Liljencrants, 1968) including several new features like a modified LF source model (Fig. 1). Then, we will describe the software environment that has been created for this kind of synthesis work. Finally, we will illustrate the approximation method by presenting two synthesized versions, using the two synthesizers, and a natural female utterance of the same sentence. Some of the remaining modelling problems will be discussed.

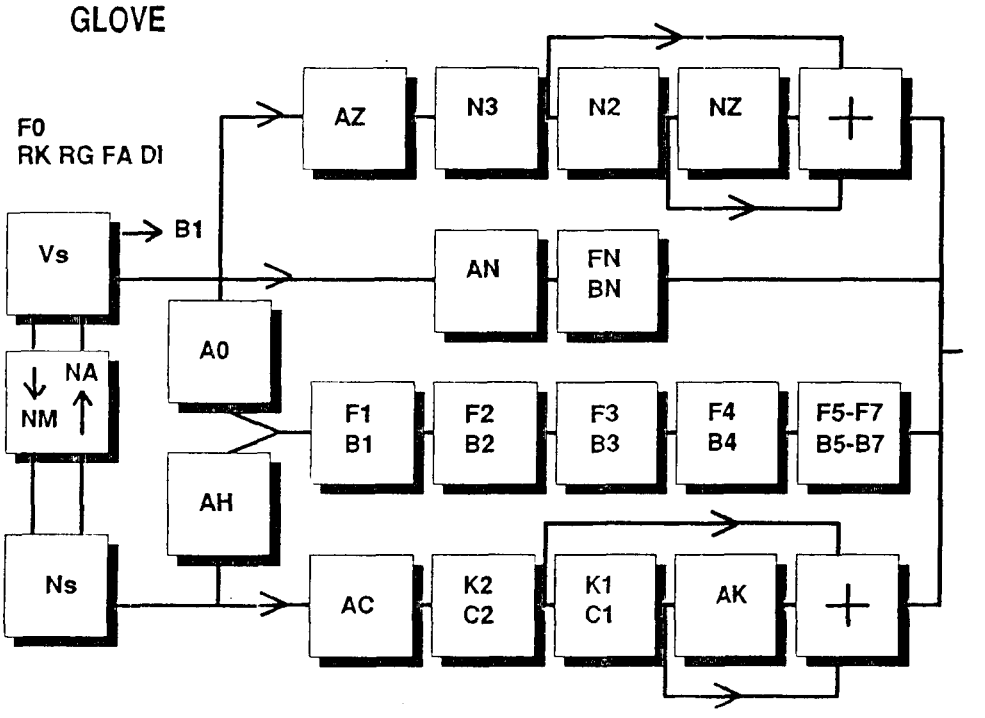


*Fig. 1. Configuration of the extended synthesizer GLOVE. For explanations of the parameters, see the text.*

## IMPROVED SYNTHESIZER

For several years we have worked with a text-to-speech system that uses a version of the now classical OVE III cascade formant synthesis model and a simplified glottal pulse source. The virtues of the simplified model have been easy parametric control and straight forward hardware implementation. The drawbacks have been the somewhat ad hoc rules to simulate the essentials of segmental glottal variations by means of bandwidth and amplitude control. We have also perceived limitations when modelling different voices. While the male voice is regarded as generally acceptable, we believe that an improved glottal source will to some extent open the way to more realistic synthesis of child and female voices as well as creating more naturalness and variations in male voices (Carlson, Fant, Gobl, Granström, Karlsson, & Lin, 1989; Karlsson 1989).

With the emphasis on signal processor realizable solutions, we have experimented with different versions of the LF-model, including storing a subset of the pulse shapes that this model generates. One version of the voice source has been the single truncated exponential sinusoid followed by variable cutoff -6dB/octave low-pass filter modelling the effect of the return phase.

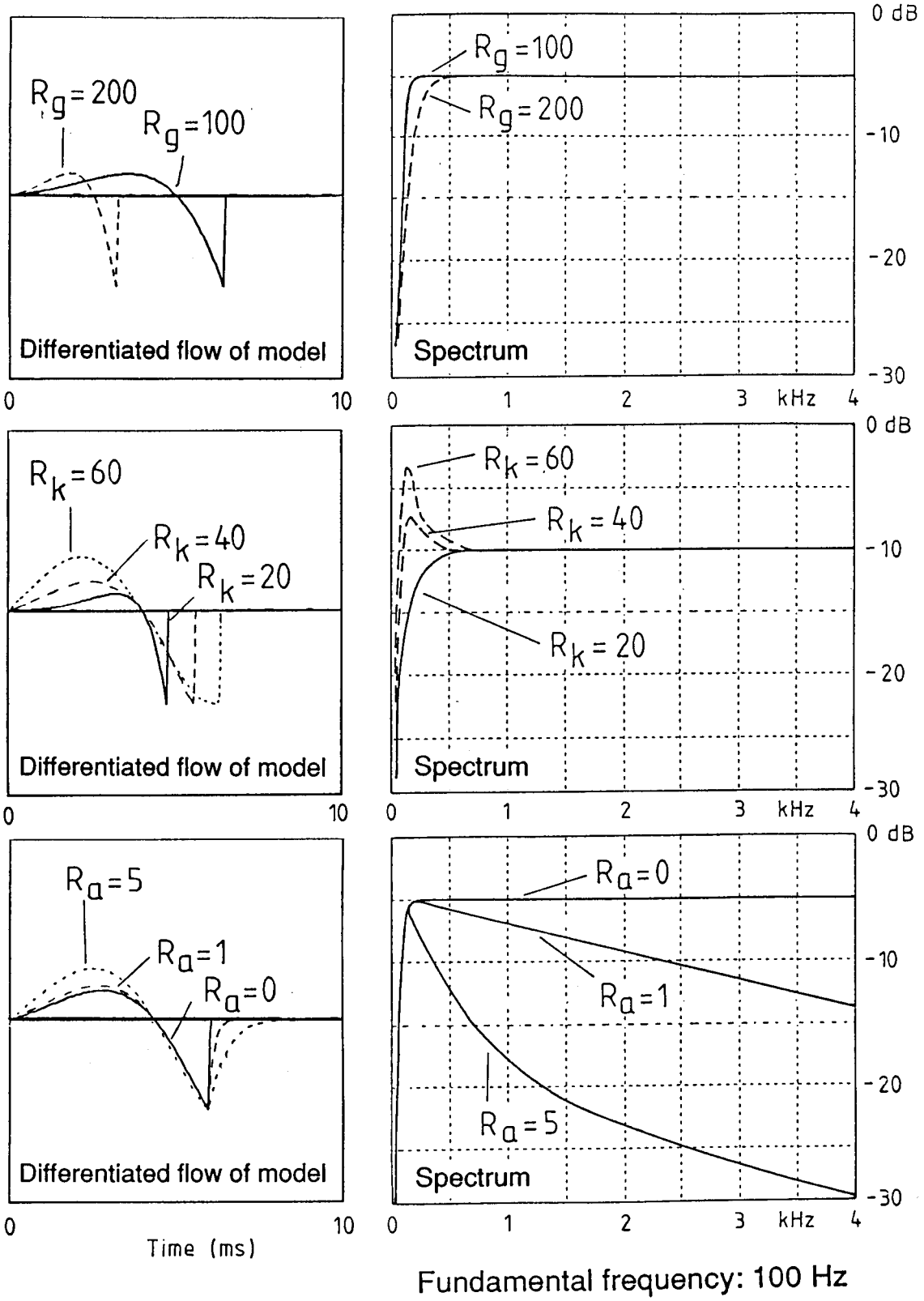


*Fig. 2.* *The influence of the parameters RG, RK and FA on the differentiated glottal flow pulse shape and spectrum. (From Gobl et al., 1989)*

This version has computational attractions but the relations to analysis parameters are to some extent compromised. The present solution is the full two-part waveform generation suggested by Fant, Liljencrants & Lin (1985). The control parameters used for this source are RK, RG, EE, FA, and F0. RK corresponds to the quotient between the time from peak flow to excitation and the time from zero to peak flow (in the differentiated flow in Fig. 2, this corresponds to the time from the zero-crossing to the negative peak divided by the time from the start to the zero-crossing). RG is the time of the glottal cycle divided by twice the time from zero to peak flow. RG and RK are expressed in per cent. EE is the excitation strength in dB and FA the frequency above which an extra -6dB/octave is added to the spectral tilt. RG and RK influence the amplitudes of the two to three lowest harmonics, FA the high frequency content of the spectrum, and EE the overall intensity. Fig. 2 explains the function of these parameters.

The synthesizer has been supplemented with a flow control of the noise source. The glottal flow waveform is used to amplitude modulate the noise. This source is used to model frication at supraglottal constrictions in the vocal tract. It is connected to the vowel branch and the parallel pole zero branch used for the synthesis of fricatives. One major advantage of the flow-modulated noise is the more convincing way to synthesize voiced fricatives. The degree of this modulation can be controlled by the parameter NM.

A new parameter NA has been introduced to control the mixing of noise into the voice source. The noise is added according to the glottal opening and is, thus, pitch synchronous as is the noise control by the NM parameter. The NA parameter models noise generated at the glottis, while NM describes supra-glottal noise generation. Either the noise source is flow modulated (NM), or noise is added to the voice source (NA). Similar methods were also used in the source model that we explored in earlier experiments (Rothenberg, Carlson, Granström, & Lindqvist-Gauffin, 1975).

Another vocal source parameter of a slightly different function is the DI parameter with which creak, laryngalization, or diplophonia can be simulated. We have adopted a strategy discussed in the paper by Klatt & Klatt (1990), where every second pulse is lowered in amplitude and shifted in time.

The "source – vocal tract" interaction is not implemented in the current synthesizer. The only control concerns the bandwidth of the first formant that is adjusted according to glottal opening. The inclusion of this feature seems to add to the naturalness of the speech in the synthesis experiments that we performed. It is, however, not clear whether this parameter has a perceptual importance compared to a simple setting of an average bandwidth. Experiments by Nord, Ananthapadmanabha, & Fant (1986) did not show a significant effect on this issue.

The synthesis of nasals have been a problem in our work to improve the segmental quality. One drawback has been the lack of extra nasal resonances. In the new model, two new poles and a zero have been added in parallel to the original nasal resonance to solve some of these problems. In a perceptual evaluation, about 2/3 of the nasal feature errors disappeared after the addition of this new branch, (Carlson, Granström, & Nord, 1990).

In our work to simulate a female voice we have found that the treatment of higher poles has to be more flexible compared to earlier models. A higher pole correction of three formants that correspond to the fifth, sixth, and seventh resonances has been introduced. The fifth is controlled like the lower formants while the following two have a fixed relation to this formant. Preliminary, these are put at a ratio of 1.2 and 1.4 relative to the fifth formant. The spectral peak corresponding to the seventh formant will not be prominent after the low-pass filtering (6.3 kHz) at the output.

One important origin of distortion in a digital synthesis model is the method to update parameters. In the old synthesizer the methods were simple. The filter coefficients were updated pitch synchronously according to the last frame from the control program (10 ms

frame update). In the new synthesizer we have tried different strategies and settled on a method to make a linear interpolation between consecutive control frames at the time for update. This interpolation is very important to avoid unwanted distortion for a high pitch voice. The parameter configuration should not be kept constant for two consecutive glottal pulses. To evaluate the importance of an even more frequent update, a sample-by-sample update of filter coefficients was implemented. No significant difference could be found.

The noise source is updated frame by frame. If the source amplitude is zero, the noise generator is reset. In this way, the same type of noise excitation is used at the release of every stop, and the place cues can be modelled more reliably.

The new synthesizer has been integrated in the control framework of the text-to-speech system. Some of the generalizations from the analysis work are now being implemented and evaluated.

## DEVELOPMENT TOOLS

The central part of the development facility is the rule development environment, where rules can be formulated, compiled, and tried out in separate rule components (Carlson, Granström, & Hunnicutt, 1990). Trace facilities of different kinds exist, from simple printouts of the results to detailed traces of individual rule applications. Statistics of rule productivity, for example, can also be gathered. On the parameter level, graphs of parameter tracks can be displayed. Support programs for maintaining frequency-ordered reference word corpora have been developed. The development environment is implemented on several computer systems, including IBM/PC compatibles, and most recently on our network of Apollo work stations. The Apollo system is also used for general speech analysis, and it has been possible to combine speech analysis and synthesis in an environment efficient for developing the phonetic component of the text-to-speech system.

One important component for synthesis development is the rule-controlled data base search system (Carlson, Granström, & Nord, 1989). The system accesses label files in our recorded speech data base. These label files have the phonetic transcriptions stored together with the time location of all segment boundaries. A search-rule system, written in the RULSYS notation, uses these files as input. Different phonemes in specific positions can, thus, be marked and stored for examination. We can, for example, formulate a rule that selects all stressed tense /i:/-vowels. These vowels can then be analyzed and presented in a contour histogram representation. The bottom graph in Fig. 3 is an illustration of this method. The synthesis model is compared to the contour histogram of 83 vowels produced by one male speaker. Since the higher pole correction is of the older type described above, it has to be placed at a rather high frequency. The higher poles will otherwise get too much emphasis.

An important complement to the rule-generated synthesis is manually controlled synthesis, so-called HI-FI synthesis or "analysis-by-synthesis". We have developed an extension of RULSYS, called HISYS, which can use rule-generated parameter data as input. This will give the user a quick start and also create a link between the two approaches. Data can then be edited according to the user's wishes. The current experiments are examples of this type of work. The phonetic transcription of a word is entered in RULSYS and the rule-controlled parameter tracks are sent to the HISYS component. The spectrogram of the same word is put under these tracks on the screen and the duration for each segment is adjusted.

With the help of a cursor, each parameter can be adjusted, points can be deleted or new points inserted. During the whole process, the synthesis can be listened to in an interactive way. The synthesized sound can also be compared to earlier versions. The user might study the spectral shape at a specific time frame in the synthesis. The point is marked and a software simulation of the synthesizer is used to create the corresponding spectrum. This can,

of course, be compared to the original recording, but as an alternative, the result of a data base search can be used.

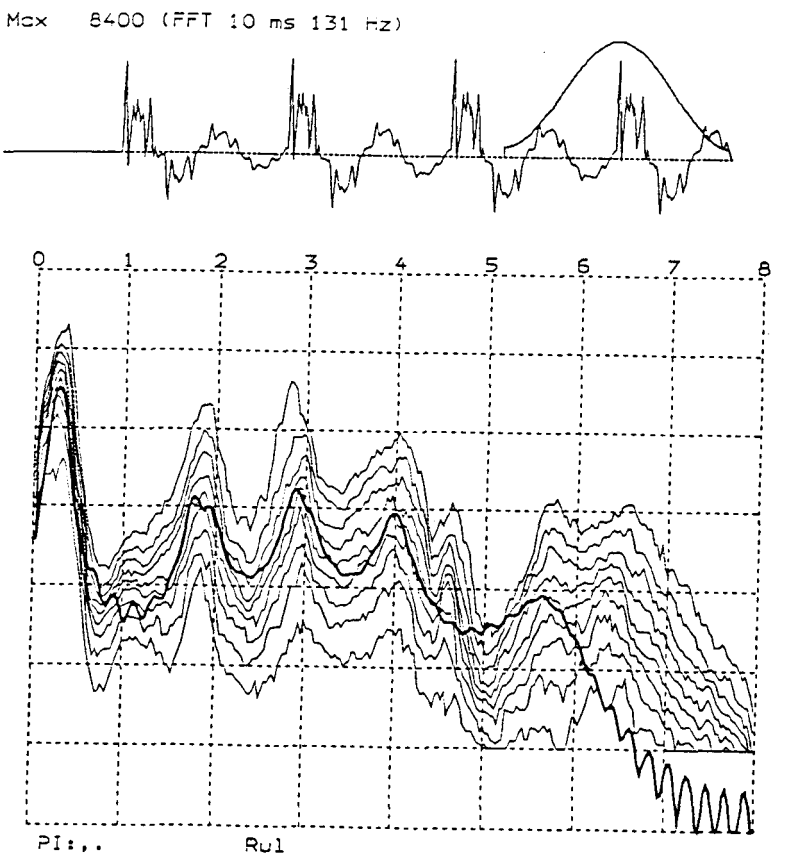


*Fig. 3.* *Upper part: Time wave and analysis window for the synthetic vowel /i:/. Lower part: Contour histogram representation of the energy distribution for 83 /i:/-vowels (thin lines) and the spectrum for the synthetic vowel /i:/ (thick line).*

## EXPERIMENTS WITH VOICE QUALITY

In order to evaluate the new synthesizer, the methods described above were used to create different voices. We wanted to produce a synthesis reference with an acceptable female voice based on a recording of a female speaker. A spectrogram of the natural Swedish sentence "Pia odlar blå violer", that was copied by synthesis, can be seen in Fig. 4. It contains a fair amount of voiced segments which is of interest to test the new glottal source.

The sentence was analyzed in terms of segment durations, formant frequencies, fundamental frequency, and glottal source parameters. In the experiment, we used a linear interpolation between turning points for the formant parameters. In the rule synthesis program, we have chosen a slightly different method where the linear smoothing is described by a target value and a time to reach this value. Both target and time can be specified by rule. This has some advantages since it is less dependent on the specific duration structure. However, in a synthesis-by-analysis experiment, the linear method is more suitable.

The measured specifications were used as control parameters for the old synthesis model to create a synthesis stimulus. A spectrogram of this stimulus is displayed in Fig. 4. As can be seen, the synthesis deviates considerably from the natural sentence. Most striking is the high degree of aspiration in the voiced segments for the natural speech which is not modelled in the synthesis.

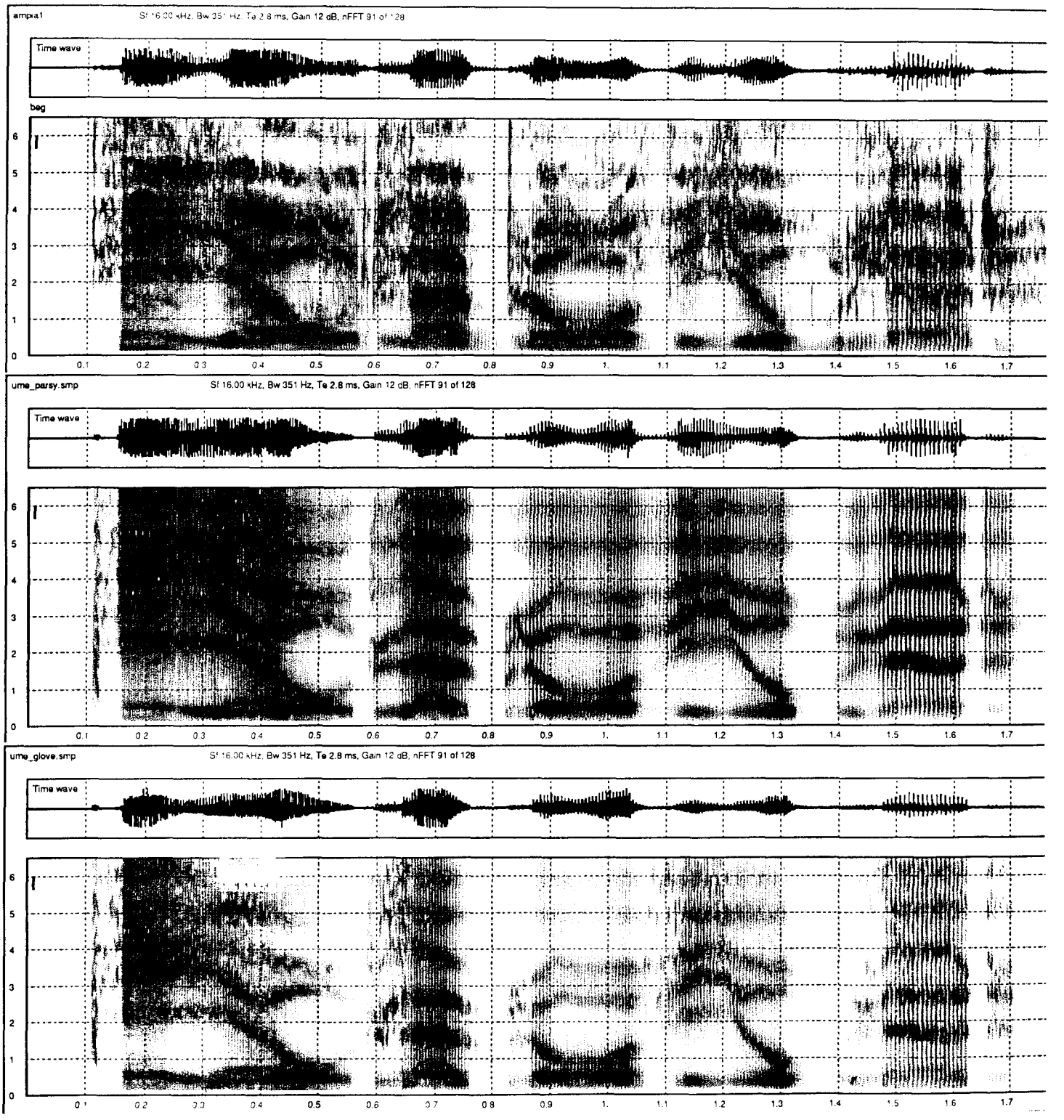


*Fig. 4.* ***Time wave and broad band spectrograms of the Swedish sentence "Pia odlar blå violer", /pi:a u:dlar blo: vlu:lər/ uttered by a female speaker (top), produced with the old synthesis model (middle) and produced with the new synthesis model (bottom).***

The sentence was inverse filtered at a few positions in each phoneme, typically at the boundary between two phonemes, in the middle of a phoneme and in the most constricted part of the long /u/-, /o/-, and /i/-vowels. A voice pulse defined by the LF-model parameters

described above was fitted to the inverse filtered pulse to get the source parameters for the synthetic voice source. An example of the analysis of the /i:/ vowel is shown in Fig. 5. (The same vowel /i:/, uttered by a different speaker is shown in our data base example, Fig. 3.)

The voice source specifications were added to the specifications used for the old synthesis model and an utterance was produced using the extended synthesizer GLOVE described above. It was found that though the quality of the synthesis was much improved, the improvement was not sufficient to achieve a good female voice quality.

Several aspects had to be modelled in addition to the formant values and the voiced source parameters. The new higher pole correction proved to be a valuable improvement for the spectral balance, but the default bandwidth of the first formant was too narrow. (Without adjusting this value, the synthesis was perceived as speech in a closed room or resonator.) We chose to use the interactive model rather than a general broadening of the bandwidth.

The major difference between our speaker and the synthesis was the breathy source. The noise had been a problem already in the inverse filtering phase. The NA parameter was used to control an addition of pitch synchronous noise. The addition of noise was specially important to mark the juncture between the /i:/ and the /a/ and in the final part of the sentence.

Addition of noise in the vocalic consonants in some contexts was also of importance. However, this type of noise is probably of a different nature and can often occur in voices with very little glottal leakage in other contexts. The frication in an /l/ sound after a stop is common but can be lacking if a morph boundary is intervening.

A spectrogram of the adjusted synthesis is shown in Fig. 4. Several aspects have been improved and the quality is significantly better than the first version of the synthesis. The segmental details can be adjusted for some of the consonants, but our major concern is the modelling of the constricted phase of the rounded vowel /u:/. The high frequency energy is still too low and can only be compensated for by adjusting the vocal source in an ad hoc way.

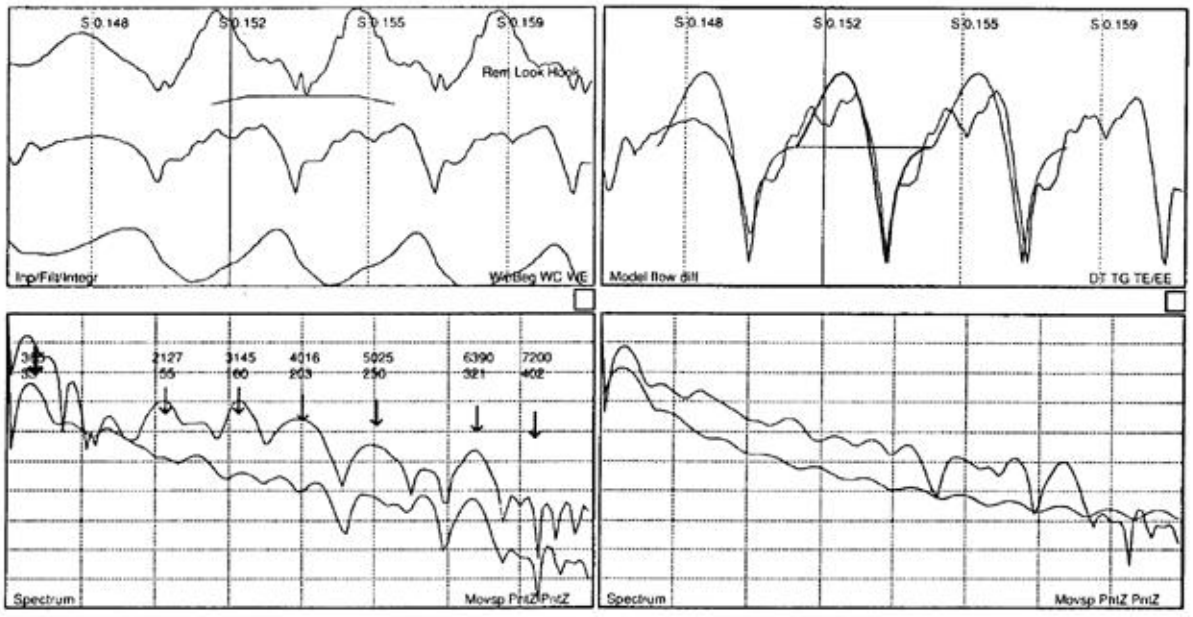


*Fig. 5. Example of inverse filter analysis. The left upper part shows from the top: the speech wave, the inverse filtered speech wave and an integrated version of it. Below this the spectrum of the speech and the inverse filtered speech is shown together with the filter settings. To the right a matching of the LF-model to a voice pulse is shown, with the time wave above and the corresponding spectra below.*

The current work has been concentrated on the modelling of a female voice. At the conference we played examples of different synthesized version of this development work. Based on this reference we also discussed possible transformations of this voice to other voice qualities.

**Acknowledgements**

This project has been supported in part by grants from the Swedish National Board for Technical Development (STU) and Swedish Telecom.

**References**

Bladon, A., Carlson, R., Granström, B., Hunnicutt, S., & Karlsson, I. (1987): "A text-to-speech system for British English, and issues of dialect and style," pp. 55-58 in (J. Laver & M.A. Jack, eds.) *European Conference on Speech Technology, Edinburgh, Vol. 1*, CEP Consultants Ltd., Edinburgh.

Carlson, R., Granström, B., & Hunnicutt, S. (1990): "Multilingual text-to-speech development and applications", in A.W. Ainsworth (ed.), *Advances in Speech, Hearing and Language Processing*, JAI Press, London, forthcoming.

Carlson, R., Granström, B., & Nord, L. (1989): "The KTH speech database," Proc. of ESCA workshop on Speech Input/Output Assessment and Speech Databases, 20-23 September 1989, Noordwijkerhout, the Netherlands.

Carlson, R., Granström, B., & Nord, L. (1990): "Evaluation and development of the KTH text-to-speech system on the segmental level," Proc. IEEE 1990 Int. Conf. on Acoustics, Speech, and Signal Processing, (21.S6a.7), Albuquerque, New Mexico, USA.

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I., & Lin, Q. (1989): "Voice source rules for text-to-speech synthesis," pp. 223-227 in *Proc. ICASSP-Glasgow, Vol.1.*

Fant, G., Kruckenberg, A., & Nord, L. (1990): "Prosodic and segmental speaker variations," pp. 106-120 in (J. Laver, M. Jack, & A. Gardiner, eds.) *Proc. ESCA workshop on Speaker Characterization in Speech Technology, Edinburgh, June, 1990*, Centre for Speech Technology research, Univ. of Edinburgh.

Fant, G., Liljencrants, J., & Lin, Q. (1985): "A four-parameter model of glottal flow," *STL-QPSR* No. 4, pp. 1-13.

Gobl, C. (1988): "Voice source dynamics in connected speech" *STL-QPSR* No. 1, pp. 123-159.

Gobl, C. & Karlsson, I. (1989): "Male and female voice source dynamics", to be published in *Proc. of Vocal Fold Physiology Conference*, Stockholm.

Karlsson, I. (1988): "Glottal waveform parameters for different speaker types," pp. 225-231 in *Proc. SPEECH'88 Book 1* (7th FASE-symposium), Institute of Acoustics, Edinburgh.

Karlsson, I. (1989): "A female voice for a text-to-speech system," pp. 349-352 in (J.P. Tibach & J.J. Mariani, eds.) *Eurospeech 89, Paris, Vol. 1*, CPC Consultants Ltd., Edinburgh.

Klatt, D. & Klatt, L. (1990): "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J.Acoust.Soc.Am.* **87**:2, pp. 820-857.

Liljencrants, J. (1968): "The OVE III Speech synthesizer," *IEEE Trans on Audio and Electroacoustics* **AU-16**:1, pp 137-140.

Nord, L., Ananthapadmanabha, T.V. & Fant, G. (1986): "Perceptual tests using an interactive source filter model and considerations for synthesis strategies," *J.Phonetics* **14**:3/4, pp. 401-404.

Pinto, N.B., Childers, D.G. & Lalwani, A.J. (1989): "Formant speech synthesis: Improving production quality," *IEEE Trans on Acoust., Speech, Signal Processing* **37**:12, pp. 1870-1887.

Rothenberg, M., Carlson, R., Granström, B., Lindqvist-Gauffin, J. (1975): "A three-parameter voice source for speech synthesis," pp. 235-243 in (G. Fant, ed.) *Speech Communication, Vol. 2*, Almqvist & Wiksell, Int., Stockholm).