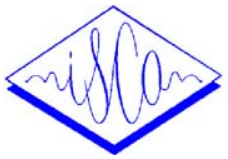

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

DESIGNING A SPEAKER ADAPTABLE FORMANT-BASED TEXT-TO-SPEECH SYSTEM

V.Darsinos, D.Galanis & G.Kokkinakis

Wire Communications Laboratory
University of Patras, 26500 Patras, Greece

ABSTRACT

First results of the efforts to build a formant Text-to-Speech system, capable to change its characteristics and imitate a specific speaker`s voice, are presented. The designing procedure is based on the automatic analysis of phonetically labelled utterances of the speaker, for the automatic extraction of formant values, voice source characteristics and coarticulation rules. All these parameters are necessary to control the synthesizer. The results of preliminary listening tests are encouraging, indicating that the system can serve as an efficient tool for the automatic analysis of speaker voice characteristics and speaker imitation.

10.21437/Eurospeech.1997-656

1. INTRODUCTION

Current speech research is focused in studying the variability of natural speech and employing the results to synthetic speech. This, because it is well understood that synthetic speech cannot be merely intelligible in order to be exploitable in many present and future applications. The potentials of speaker and speaking style variations should be incorporated in advanced Text-to-Speech (TTS) systems. For example, in future translating systems, ideally the speaker characteristics should be preserved in the translation[1].

The current work in voice personalization and speaker adaptation is concentrated in the area of speaker transformation using time-domain techniques based on unit-concatenation speech synthesis schemes[2][3]. The main idea behind these methods, is to try to reflect the characteristics of one speaker into another using vector parameters and rules. Unfortunately, there are serious difficulties in building and controlling a unit-concatenation system (optimal unit selection, coarticulation phenomena, signal continuity etc.). Furthermore, the system´s database needs to be rebuild for each speaker in order to be able to produce acceptable speech of unlimited vocabulary.

In this work we describe a different approach which uses frequency domain parameters and ensures quick speaker adaptation. The whole procedure is based on the automatic analysis of phonetically labelled utterances of the speaker which provides all the necessary speaker-dependent parameters to control the formant synthesizer.

This procedure has been tested on a TTS-System for Greek and has given encouraging results.

2. SYSTEM DESCRIPTION

The designed system consists of three layers: 1) the speech analysis module, 2) the parameter processor and 3) the formant TTS-synthesizer.

A prerecorded and prelabelled set of speaker speech signals is analysed in the first and second layer in order to extract the vocal tract parameters, the voice source parameters, the mean phoneme duration and the mean phoneme energy. The third layer consists of a phoneme-based pole-zero formant synthesizer, developed in our laboratory, which is able to produce speech of good quality and high intelligibility. Figure 1 presents the complete system.

2.1. Extracting speaker parameters

Speaker signal analysis consists of calculating the vocal tract and excitation parameters of the speaker. All the analysis procedure is carried-out only on the prerecorded speech signal. In this way speaker adaptation is achieved with the less effort and information possible.

Vocal tract parameters are estimated from the closed-phase of the signal. The points of the glottal closures are marked through an epoch detection procedure based on the cross-correlation of the signal and LPC-wavelets[4]. This procedure is a modified version of the epoch extraction method proposed in [5]. The modified version uses the maximum-likelihood epoch determination to locate the glottal closure points and a linear detector to improve performance and reliability. In Figure 2 the structure of the algorithm can be seen. According to this method, using a shifted window, the interval between two successive points is

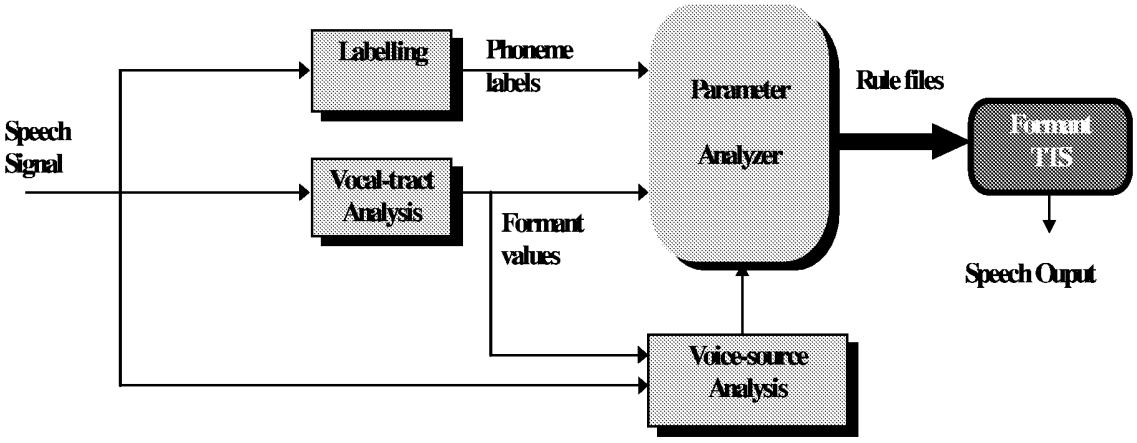

Figure 1. System´s description

searched, in order to find a residual error minimum that determines the closed-phase areas. The complete method is described in [4].

The last step consists of modelling, using the LPC-covariance method, the optimally positioned windows, that are supposed to lie in the closed-phase interval. The LPC polynomial is solved to estimate the vocal tract parameters.

The use of a N-best formant tracking algorithm [6] ensures the continuity of the estimated parameter tracks.

We adopted a tracking-algorithm due to the difficulties to estimate the formants when the short time spectrum is relatively flat or ambiguous. In this case the location of the formants can only be determined by tracking the formants. In all other cases the LPC analysis can be done successfully.

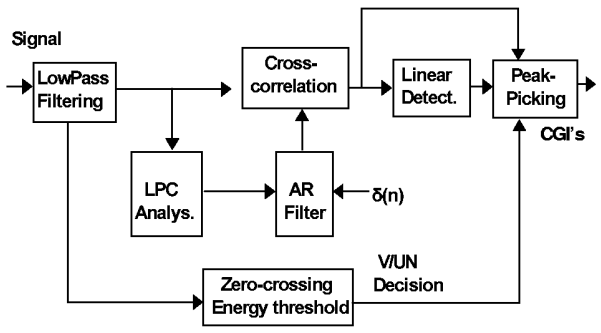

Figure 2. The Glottal Closure detection algorithm

In addition, a tracking algorithm must be allowed to compensate for incomplete local information. It has been observed that for most speech segments, only a few consistent interpretations of formants can be made. Thus, designing a formant tracker that finds the N-best tracks seems to be the right solution, instead of a single best tracker [7]. In this case if the estimation of a formant is obscured, the missing information can be recovered by using consistency constraints with respect to the adjacent frames. Thus we reduce the problems appearing in some heavily glottalized speech or speech distorted by several noise events.

The algorithm used in this work consists of three steps. First, the elementary formant tracks are estimated connecting the formant candidates and using frequency constraints for each formant and constraints of the types F1-F2 and F2-F3. Then a least-square polynomial method determines the formant regions and their distribution and assigns each candidate to an elementary track. Using the elementary tracks as they are computed from the above procedure, a formant selection, correction and connection of individual tracks must be done. To this end least-square 3rd order polynomial tracks are computed for each formant, using the elementary tracks. After that, each candidate formant quartet is examined using distance criteria.

The final processing step reconstructs the lost or empty regions of the formant tracks using cubic spline interpolation. This procedure guarantees a formant smoothness and does not introduces any artifacts [6].

Voice source parameters are estimated pitch synchronously through model-fitting on the inverse filtered signal. The Liljencrants-Fant (LF) model of the source is used because it is realistic and efficient. In addition its control is less complex as it is determined by only four parameters.

The estimated glottal closure points are used for a reliable definition of the periods. The LF-Model parameters are estimated in an iterative way minimising the quadratic error between the estimated parametric model and the measured signal.

In this procedure we reduce the problems related to least-square optimization techniques, like the appropriate selection of the initial estimates and falling into local minima without convergence. Furthermore, it gives an acceptable estimation in all cases, against an increase of the computational cost. The extracted parameters are stored and processed in order to control the voice source model of the synthesizer.

2.2. The parameter processor

The function of the parameter processor is to create the necessary rule files, using the stored results of the analysis procedure and the corresponding labelled speech signals, in order to feed the TTS-System. The TTS-system used in this work employs a pole-zero phoneme-based synthesizer excited by the parametric LF-Model. The coarticulation, intonation and segmentation effects are modelled by the use of special text rule files that are compiled to source code, following the symbolism used in the ESPRIT POLYGLOT project. According to this scheme the individual phoneme parameters can be modified depending on the phonetic context. Special characters are used to show the different types of phonemes that are considered.

The parameter processor provides two sets of data:

1. Text files with the phoneme parameters along with their target values (formant values, duration, duration of formant transitions). These files can be used directly by the TTS-System for speech generation.
2. Files with the appropiate rules simulating the coarticulation phenomena and modelling the deviations and transitions from the target values. These files are converted to source code and linked with the main synthesizer.

It is important to notice that the reliable and efficient determination of coarticulation rules, depends on the accuracy of the labelling of the phoneme transitions. For this reason a manual labelling procedure was applied since we wanted to reduce the influence, on the final speech quality of problems concerning the analysis procedure and concentrate in the examination of the efficiency of the proposed system.

In this context the labelling of the speech signals was carried out using special software developed in our lab that permits the user to see the signals in both domains, time and spectral, and listen to it. Our aim is to automate this procedure by exploiting a phoneme segmentation technique developed and used in speech recognition[6].

3. VALIDATION

To validate the methodology used in imitating a specific speaker's voice and evaluate its efficiency, we performed preliminary tests with 5 male speakers and 6 listeners. In order to eliminate the influence of the quality of the synthesizer, all the listeners were chosen to be familiar with synthetic speech.

Two sets of speech signals were used as stimuli: 6 sentences (set 1) and 18 VCV and CVC syllables (set 2), recorded by each speaker.

As training corpus a set of 25 meaningful sentences different from those used as stimuli, which were taken from newpapers and contained at least once all the phonemes, was recorded by each speaker. The speakers were asked to read aloud the given text in a neutral way. This speech material was labelled and analysed with the procedure described and the results were used to create the synthetic versions of the test signals.

The natural and the synthetic versions of the signals were then presented to the listeners who were asked to identify the speaker in each one, after listening to a speaker speech demo. The listeners were told not to concentrate on the meaning of the speech but to the voice and speaker characteristics. The total mean results of the above procedure for each set of signal stimuli, are presented in Table 1.

	Set1		Set2		**Mean**	
Listeners	Natural	Synthetic	Natural	Synthetic	**Natural**	**Synthetic**
1	**100**	**80**	**62**	**62**	**81**	**71**
2	**100**	**60**	**68**	**56**	**84**	**58**
3	**80**	**80**	**87**	**56**	**83.5**	**68**
4	**100**	**80**	**62**	**50**	**81**	**65**
5	**60**	**60**	**56**	**37**	**58**	**48.5**
6	**95**	**72**	**80**	**52**	**87.5**	**62**

Table 1. Success rate of speaker identification (%)

The results showed that in the case of the non-sense syllables the listeners have a difficulty to identify the speaker, even from the natural speech. In contrast, in the case of the first set the identification is more successful. These first results are encouraging in the sense that the method used for extracting the speaker characteristics is efficient and can serve as the basis in speaker speech analysis. On the other hand additional work must be carried out in the automatic extraction of the articulation rules, especially for some consonants and plosives, in order to improve the segmental quality of the produced speech.

4. REFERENCES

[1] B. Granström, ¨The use of speech synthesis in exploring different speaking styles¨, STL-QPSR 2-3, 1991.

[2] H. Valbret, H. Moulines & J.P. Tubach, ¨Voice transformation using PSOLA technique¨, Speech Communication Vol.11, pp.175-187.

[3] K. Lee, D. Young & I. Cha, ¨Voice personality transformation using an orthogonal vector space conversion¨, Proc. EUROSPEECH 95, pp.427-430, Madrid 1995.

[4] V. Darsinos, D. Galanis & G. Kokkinakis, ¨A method for fully automatic analysis and modelling of voice source characteristics¨, Proc. EUROSPEECH 95, Madrid 1995.

[5] Y. Cheng & D. O´Shaughnessy, ¨Automatic and reliable estimation of glottal closure instants and period¨, IEEE Tr.on ASSP, vol.37 (12), 1989.

[6] J. Sirigos, V. Darsinos, N. Fakotakis & G. Kokkinakis, ¨Vowel/Non Vowel classification of speech using an MLP and rules¨, Proc. EUSIPCO 96, Trieste Italy 1996.

[7] Y. Laprie, M. Berger, ¨A new paradigm for reliable automatic formant tracking¨, Proc. ICASSP 94. pp.345-348, 1994