FLEXVOICE: A PARAMETRIC APPROACH TO HIGH-QUALITY SPEECH SYNTHESIS

György Balogh, Ervin Dobler, Tamás Grőbler, Béla Smodics, Csaba Szepesvári *

ABSTRACT

The TTS system described in this paper is based on the analysis and resynthesis of a given speaker's voice. First, the speaker's voice definition is prepared off-line: a diphone database is recorded, segmented, and analyzed in every 6 msec to obtain the filter parameters of an all-pole (AR) filter. During the on-line synthesis, the filters are excited with the mixture of a predefined periodic glottal source and white noise. Rigorous experiments have been made to find the parameter space in which the filter coefficients at diphone boundaries can effectively be smoothened. The best representation turned out to be the space of area ratios. Due to the smoothening and the carefully chosen corpus words, each diphone needs to be recorded only once thus no unit selection algorithm is needed. FlexVoice provides large flexibility in changing voice properties independently from the vocal tract parameters. This flexibility can be demonstrated by a number of voice conversions including female-to-male and female-to-child conversions. FlexVoice only uses a fraction of the resources of a PC and its quality is comparable to that of the leading TTS systems.

1 INTRODUCTION

The majority of recent achievements in high-quality text-to-speech (TTS) synthesis have resulted from time-domain concatenative synthesis. The alternative model-based (or parametric) approaches have often been judged less powerful because they provide inferior segmental quality (Dutoit 1997). Wave concatenation, however, also has obvious shortcomings that can only be overcome by using a model-based approach and trying to achieve the segmental quality of time-domain methods.

An early but still prominent parametric model is the source-filter model (Fant 1960) that treats the glottal source separately and views the vocal tract as a filter acting on this source. A typical family of filter models is the autoregressive (AR) model that represents the vocal tract with a very limited number of parameters. Though intelligible speech can be synthesized by rule-based generation of the filter parameters (e.g. Klatt and Klatt 1990), natural-sounding synthesis can only be achieved by the analysis of human voice. Together with a parametric description of the glottal source, this approach yields the following advantages over time-domain methods:

- Easy prosody matching
- Smaller database
- Simple concatenation method
- An easy way of voice manipulation and conversion

Despite these advantages, parametric synthesis has not been able to compete with the current high-quality TTS systems.

FlexVoice is an integrated text-to-speech technology that attempts to produce high-quality natural speech using the parametric approach. The current technology has emerged from a speech analysis–resynthesis system (called IFS) that uses the same filter analysis, source estimation, and synthesis algorithms to resynthesize any given speech fragment of any speaker. FlexVoice encompasses the linguistic preprocessing modules that are indispensable for high-quality speech synthesis but their discussion is beyond the scope of

* Mindmaker Ltd., Budapest, Hungary
Email: grobler@mindmaker.hu

this paper. The method discussed here is inherently language-independent though it has only been realized for US English.

FlexVoice provides high flexibility in modifying, converting voices, switching between voices, and singing. The following sections describe the synthesis algorithm systematically. The major achievement of this work is the careful design of each step since they all have significant influence on the overall quality of the synthesis.

2 SPEAKER SELECTION

It is very important to select the speakers carefully because all of the following steps may suffer from an inappropriate selection. Both the speakers' linguistic abilities and their voice properties have been rigorously examined. To meet the requirements, the speaker should

- be native speaker of the language (US English)
- have a "standard" dialect that is acceptable for the language community
- be able to articulate clearly and naturally at the same time
- have a pleasant voice when synthesized

To decide whether any speaker fulfills the above requirements, a simple test was designed. Seven female and six male speakers were tested. A test sentence was selected that contained all types of phonemes that might be problematic during synthesis. Approximately 50 diphones, sufficient to synthesize the sentence, were collected. Speakers were asked to read the small test corpus consisting of sample words that contained the diphones. No sample word was allowed to be the same as in the test sentence. The same test sentence was then synthesized in all speakers' voices.

Five linguists were asked to judge whether the speakers' pronunciations were appropriate. Though the most striking accents had been rejected during the telephone conversations before the test, still one female speaker was excluded because of her dialect.

Ten subjects were asked to evaluate the quality of the synthesized voices by listening to the synthesized versions of the test sentence only. To compensate different rating strategies, subjects had to divide 100 points among the different voices. Finally, two female and two male voices were selected.

3 THE DIPHONE DATABASE

One of the key points of the technology is the design of the diphone database. This is even more so because all diphones are represented with a single word (or phrase) that is recorded only once.

First, the set of phones to be used should be determined. The phone system used in FlexVoice contains 56 sounds. The English version of the SAMPA system has been modified and adapted to US English, then allophonic variations such as aspirated stops, syllabic consonants, diphthongs, etc. have been added.

While the phonemes can in principle occur in any context, the allophone generation rules exclude certain configurations of phones. With this in mind, occurrences of all possible diphones have been looked for. In some cases, the diphones cannot be found in single words. In such cases, word pairs containing the diphones at the word boundary have been selected. A number of additional constraints arise from the requirement that the same sounds in different diphones should be uniform all over the database so as they can be matched during synthesis. Some constraints of this type are the following:

- Diphones should not be at the beginning or end of words
- Vowels that can be stressed are taken from stressed syllables
- Vowels should possibly not be followed by nasals or liquids
- Voiced stops and fricatives should not be followed by voiceless ones

The requirements can be fulfilled by careful selection of words and, where necessary, application of phrases that contain the target word(s). A total number of about 2200 words/phrases have been collected. The remaining 400 diphones are theoretically possible but hardly ever appear in English speech.

4 RECORDING

It is quite evident that the quality of the recorded speech is crucial. Some factors, however, have a strong effect while others do not. A striking example is that the analysis-synthesis system has been found quite robust against background noise. The method is also robust against pitch fluctuations. Nevertheless, high-quality studio recording is indispensable for the following reasons:

- Spectral properties of the voice should be preserved with high accuracy
- Reasonable dynamic range should be provided
- Loudness should be kept constant
- Side-effects of speech such as pops should be filtered out

Studio recordings thus have been made in CD quality (44.1 kHz, 16 bit) and then downsampled to 16 kHz. Recording the diphone database with a single speaker takes approx. 5 hours.

5 SEGMENTATION

The role of accurate segmentation cannot be overemphasized since the quality of synthesized speech can be completely ruined by segmentation errors. Unfortunately, no automatic segmentation algorithm has been found sufficiently accurate; thus, segmentation is currently done semiautomatically. Nevertheless, a segmentation tool has been developed to display the wave files and their spectra visually and help the expert place the markers.

Three markers are used to mark a diphone:

1. inside the first phone
2. at the border of the two phones
3. inside the second phone

The exact positions of the first and last markers inside the phones are determined by the phonetic properties of the given sound, still considerable expertise is needed to place the markers correctly.

6 VOICE DEFINITION

Once the recorded wave files are segmented, voice analysis can be performed and yield the parameterized information describing the given speaker's voice. This information is called voice definition and consists of the following parts.

6.1 Diphone parameters

The most important part of the voice definition is the parametric description of the diphones in the database. It should be noted that only one instance of each diphone is segmented. This is the only information about the diphones that is available for the synthesis algorithm.

6.1.1 Parameter packets

The speech waveform inside each diphone target is windowed and analyzed and the relevant parameters are stored in a parameter packet. During synthesis, the parameter values in a packet are valid until the values of a new packet are set. The packet contains information about its position within the diphone and the following parameters.

6.1.2 Spectral analysis

The vocal tract model of FlexVoice belongs to the family of autoregressive (AR) models. The parameters of the all-pole filter are determined with the Levinson–Durbin algorithm. Since experiments have shown no significant impact of pitch-synchronous analysis, Gauss-windows with constant window shift (6 ms) are used.

Each packet contains 20 filter parameters (prediction coefficients) and a normalization factor. Filter normalization is critical because the loudness values of different diphones must be equalized. The pitch-

dependence of the normalization factor is ignored; filter normalization can thus be an off-line process. The easiest way of estimating the normalization factor is then by measuring the energy of the synthesized wave.

6.1.3 Loudness

A target loudness value is assigned to each phone in order to equalize the loudness of synthesized speech. Target loudness values are supposed to be valid in the middle of each phone, i.e. at the diphone boundaries. The time-dependent changes of loudness inside the diphones are preserved and adjusted to the target values. The target values are computed from statistics of loudness throughout the segmented diphone database.

6.1.4 Source estimation

It is assumed that the AR filter is excited with a mixture of two sources: a periodical source as described in Sect 6.2 and white noise. The goal of source estimation is to determine how much of each source the analyzed wave contains.

The total loudness of each packet is distributed between the amplitudes of the glottal source and the noise source. The proportion given to the glottal source is called periodicity. The value of periodicity is computed as the square root of the ratio of the first and the zeroth autocorrelation peaks. Finally, both loudness and periodicity values are conveniently adjusted by phonetically inspired rules.

6.2 Glottal source

A natural way of modeling the glottal source would be to analyze the signal remaining after inverse filtering by the vocal tract filter. Experiments have been made to describe the glottal source waveform by a radial basis function neural network but results are not convincing. Thus, for the moment, the glottal source waveform is not analyzed but a parametric waveform is used instead. Finding a more appropriate glottal source model is still an open research problem.

The current version uses the KLGLOTT88 glottal source proposed by Klatt and Klatt (1990). Default values of the glottal parameters (such as base pitch, open quotient, breathiness, and spectral tilt) are set for each speaker in the voice definition. They are normally kept constant throughout the synthesis but can also be modified to manipulate the voice properties.

7 ON-LINE SYNTHESIS

The previous sections described the off-line preparation of the voice definition of a given speaker. Now the voice definitions of the speakers should be used to efficiently synthesize high-quality speech. The main steps of the on-line synthesis are the following:

1. The linguistic modules of FlexVoice prepare the phoneme string and the prosody information (duration and pitch values) for each phoneme.
2. For each diphone in the phoneme string, the corresponding parameter packets are selected from the voice definition.
3. The packets are modified to match the actual prosody.
4. The packets are further modified by segment concatenation.
5. The synthesizer generates the wave with the resulting control parameters.

7.1 The synthesizer

The main modules of the synthesizer are shown in Fig 1. Each control parameter packet is valid until the next packet is set, i.e. no interpolation of the parameters is necessary in the synthesizer. The control parameters and their target modules are listed below:

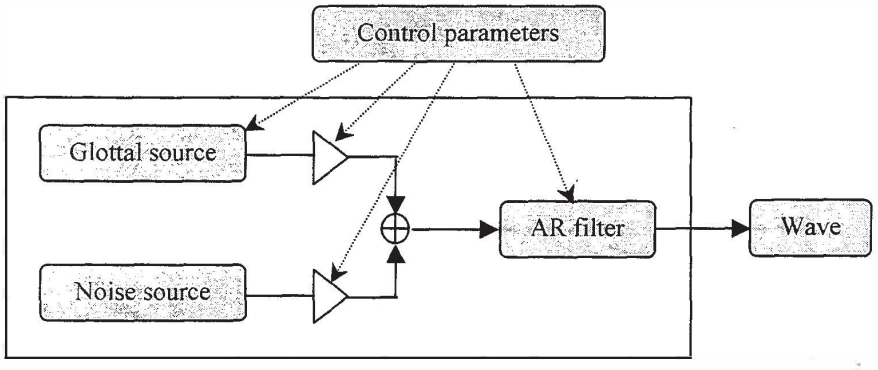

Figure 1: Schematic diagram of the FlexVoice synthesizer.

Parameter	Target module
actual pitch	glottal source
amplitude of voicing	glottal source
amplitude of noise	noise source
prediction coefficients	AR filter

7.2 Prosody matching

Prosody matching is fairly simple in FlexVoice. Since the glottal source is separated from the vocal tract filter, pitch can directly set the fundamental frequency of the glottal source. The current prosody generator provides pitch values at phoneme borders; pitch is thus linearly interpolated inside the phonemes.

Phoneme durations proportionally modify the frame lengths of the parameter packets, i.e. the number of packets is constant. The packet frame rate thus follows the tempo of speech, more packets corresponding to fast changes.

7.3 Segment concatenation

One of the major difficulties of concatenative synthesis is to eliminate discontinuities at the segment borders without introducing artifacts in the synthesized speech. In FlexVoice, this problem can be solved by interpolating the control parameters listed in Sect 7.1.

The source amplification factors can easily be smoothened by simple linear interpolation. Prediction coefficients, however, should not be interpolated because the stability of the corresponding filters cannot be guaranteed. Nevertheless, prediction coefficients can easily be transformed into equivalent representations that preserve stability when interpolated. Such representations are provided by PARCOR coefficients, line spectral frequencies (LSF), area ratios, and log area ratios (LAR). Experiments with interpolation in these representations have shown that best results can be obtained using area ratios (see also Dutoit 1997, p. 215). Thus, prediction coefficients are transformed into area ratios and linearly interpolated at the segment boundaries.

8 TESTING

Many of the above steps are prone to errors therefore high-quality synthetic speech can only be obtained after exhaustive testing. Listening to the synthesized version of the full diphone database helps to find recording and segmentation errors. Some of the errors, however, only occur when particular diphones are concatenated. Such errors can be found by extensive listening tests.

9 VOICE CONVERSION

One of the advantages of parametric synthesis over other methods is its high flexibility in manipulating the voice properties of a speaker. FlexVoice makes use of this advantage by providing the possibility of on-line modification of several voice features. The following parameters can be used to modify the speaker's voice:

- default pitch
- minimum pitch
- maximum pitch
- intonation level
- volume
- breathiness (relative amount of noise)
- head size (shift of vocal tract transfer function)
- creakiness (quick random pitch modulation)
- richness (glottal source open quotient)

Combining these parameters, one can both make voices with strange effects and convert a voice to other natural-sounding voices. Conversions from female to male and from female to child voices have successfully been made.

10 CONCLUSIONS

Subjective listening tests have shown that FlexVoice can produce high-quality synthesized speech that compares favorably with competing products. FlexVoice only needs a fraction of memory and processor capacity available in nowadays' personal computers. The current version, far from being optimal, uses 2 MB RAM per base voice and about 15 percent of CPU time of a P III / 450 MHz processor. In addition, storage of a single voice provides a number of significantly different manipulated voices.

Some steps of the technology can still be largely improved. Future work includes, among others, the development of automatic segmentation and improvement of the glottal source model.

REFERENCES

Dutoit, T. (1997): *An Introduction to Text-To-Speech Synthesis*. Kluwer Acad. Publ., Dordrecht.

Fant, G. (1960): *Acoustic Theory of Speech Production*. Mouton, The Hague.

Klatt, D.H. and Klatt, L.C. (1990): Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. *J Acoust Soc Am* **87** (820–857).