A SYNTHESIS METHOD BASED ON CONCATENATION OF DEMISYLLABLES AND A RESIDUAL EXCITED VOCAL TRACT MODEL

Steve Pearson, Nicholas Kibre, Nancy Niedzielski

Panasonic Technologies, Inc. / Speech Technology Lab
Santa Barbara, CA 93105

ABSTRACT

This paper describes the back-end of a new, flexible, high-quality TTS system. Preliminary results have demonstrated a highly natural and intelligible output. Although the system follows some standard methodologies, such as concatenation, we have introduced a number of novel features and a combination of techniques that make our system unique. We will describe in detail many of the design decisions and compare them with other known systems. A demonstration of the speech quality with implanted prosody is available in waveform file ([WAVE stltts1.wav and stltts2.wav]) on the conference CD.

1. INTRODUCTION

Modern Text-to-Speech (TTS) systems achieve a high level of intelligibility, yet for many applications, a lack of naturalness hinders their acceptance. This unnaturalness is most often attributed to the prosody generation and/or voice quality. The present paper addresses the second of these problems.

Although traditional formant synthesis [1] has many positive features, extensive research efforts in this area have not managed to yield a high level of naturalness [2,3]. Physical modeling is another compelling methodology, however the complexities involved seem to put its practical application in TTS systems out of reach in the near future. Like many others, we have decided on a concatenation-based method. Concatenation is computationally feasible and allows for a significant increase in naturalness.

It has been observed that even a small glitch or anomaly during synthesis can sometimes throw off the intelligibility of an entire sentence. So in some sense, a synthesizer is no better than the worse case concatenation units in the worse case contexts [4]. Hence, rather than a corpus based approach, we have aimed at developing concatenation unit sets with complete coverage of the language, and such that the worse case unit surpasses a certain minimum size and quality. As described in more detail later, we have chosen the demisyllable as the typical concatenation piece.

A separate problem is that of digitally representing the concatenation units. The speech data must be encoded in such a way that (1) the resynthesis is computationally efficient, (2) the memory space required is manageable, and (3) prosodic modification is feasible and induces a minimum of distortion. Many methods have been proposed: TD-PSOLA [5], harmonic/stochastic [6], LMA [7], LPC, to mention a few. We will describe our investigation of a residual-excited all pole vocal tract model.

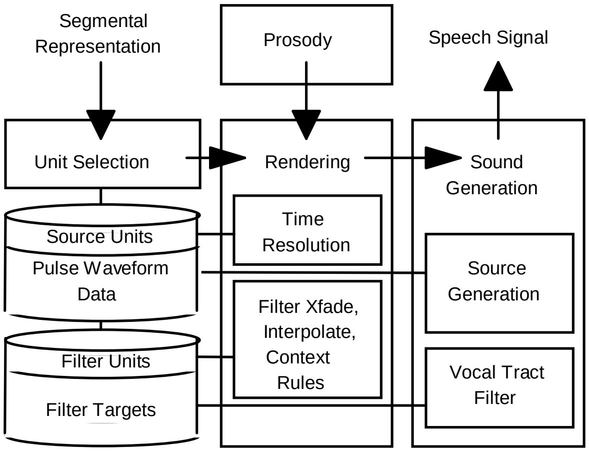

Figure 1: TTS system back-end architecture.

We aim to model as closely as possible the natural separation between the glottal source and vocal tract. We believe that this gives an advantageous balance in complexity between the source and filter in the source-filter paradigm. In what follows, more specific details of the advantages of this source-filter model will be discussed, and our method will be contrasted with other TTS specific coding methods.

We will describe the back-end portion of a TTS system that is currently being developed (fig. 1). The input is the phonetic string to be spoken and appropriate prosodic control parameters; the output is a speech waveform. Our goal is highly natural and intelligible TTS synthesis which imitates the qualities of our model speaker. In addition, we expect to eventually implement multiple voices and languages, hence we require a development methodology which is highly automated. We will discuss features of our development environment. Finally, we need a scalable system which can be adjusted according to tradeoffs between quality, size, and computational requirements. We describe how our system will be able to achieve this.

2. CONCATENATION UNIT DATABASE

Choosing the syllable as a concatenation unit leads to many pieces (more than can be handled in the present project), however by collecting demisyllables (half syllables) we can form run-time syllables from static sets of initial half syllables (onset and nucleus) and final half syllables (nucleus and coda), and thus maintain the syllable concept. It was felt that an important reason to consider the syllable unit is that often the coarticulation effect is lower at the syllable boundary than other choices. Also, as the syllable is a key point of overlap between the organization of segmental phonology and prosody, it is an optimal point to interface the prosody and sound generation modules of a synthesis system. In part, we chose to build the syllable out of demisyllables since they permit us to cross-fade in the steady-state portion of the vowel, which allows for seamless transitions.

The run-time syllables must also be joined together, and at least in English this is often very critical and difficult. Hence we have developed the complimentary concept of a run-time boundary unit that joins syllables. The boundary unit is intended to capture any features that cue a syllable or word boundary. In general, the boundary is a consonant cluster, which must also take account of the vowel types in the surrounding syllables. In many cases the boundary unit includes speech waveforms, but in some cases, may just be rules telling how to join the two syllables. In order to reduce memory requirement, the waveform associated with the boundary unit is made up of one or more diphones, but in some cases, larger groups of phonemes are needed.

An example of a boundary decomposition is the word boundary in "least school". The final demisyllable before the boundary is the /iys/ from a word like "peace". The initial demisyllable following the boundary is the /kuw/ from a word like "scoot". The key boundary piece is /ts/ across word boundary. This may already exist from a simpler boundary unit which reflects the vowel context such as for "eat soon", but it must be ascertained which is an appropriate /ts/. The /ts/ piece will be connected on the right by cross-fading in the /s/ with another /s/ from an initial /sk/ cluster, such as from "scoot". Such consonants from initial clusters have to reflect somewhat the effect of the following vowel, so we will have at least three classes, like from "skeet", "skut", and "scoot". The /s/ to /k/ connection is in the silence, and is only required to produce the proper silence duration. An additional /s/ piece from a word like "feast" must be spliced via cross-fading with the /s/ in /iys/. Again we would expect to have at least 3 such /s/'s from "feast", "gust", "boost". So this boundary unit is broken into possibly 3 pieces of size diphone or less.

Our full waveform database targets approximately 2000 demisyllable units, and 2000 boundary-piece units. The demisyllable units represent all the pieces needed to create every possible English syllable at both primary and secondary levels of stress. In addition, we have included the pieces necessary to create every unstressed syllable in the English language. The boundaries include both the high stress to low stress and low stress to high stress transitions. In the highest quality synthesizer version, we will use all the data, but in a scaled down version, statistical and clustering techniques can be used to eliminate redundant or rare data.

To create the database, we recorded approximately 6000 sentences of random style which contained within them the target demisyllable or boundary piece. The sentences were read by a professionally-trained female speaker in 400 sentence increments, and the recording was carefully monitored to assure consistency in terms of rate of speech, pitch, and intonation.

The recording took place in a deadened sound proof booth. The sentences were digitized at 11025 Hz into a computerized database. After automatic HMM segmentation, the target segments were cut from the sentences, and labeled according to segmental information.

3. UNIT PARAMETRIZATION AND SYNTHESIS MODEL

We experimented with several methods of acoustic representation for concatenation units: (1) TD-PSOLA, (2) LPC, (3) harmonic synthesis, and (4) a cascade of second order formant modeling filters with a residual-based voice source. Harmonic synthesis was rejected since the high computational requirement would be prohibitive in certain applications. The other methods had more subtle drawbacks.

TD-PSOLA exhibited a scratchy distortion with moderate pitch alteration. This is probably due to inaccuracies in pitch epoch marking, but also maybe due to phase mismatch between overlapping pulses. Also, since spectral modification is difficult, a large cross-fade at concatenation boundaries caused blurring of the formants, while a short cross-fade often caused a spectral discontinuity.

Many glitches or warbles were encountered with LPC resynthesis. Presumably, this is caused by reassignment of poles to new resonances, causing discontinuity. Also, since LPC models the source as well as the vocal tract, a greater chance of distortion during parameter interpolation is expected.

For a number of reasons we have chosen to currently focus on method (4). First, the voice source which is derived by inverse filtering the speech data is devoid of major vocal tract resonances and hence avoids harmonic phase mismatches in prosodic modification and concatenation. Secondly, formant frequencies and bandwidths can be easily interpolated and in so doing formant blurring can be avoided during a large cross-fade.

Also useful in this case is the vast body of knowledge linguists have developed in terms of formants. A parameterization in terms of formants allows the use of this knowledge base in formant modifying context rules. For example, vowels are known to be somewhat neutralized in fast speech, and this could be captured by a rule which moved formant parameters towards neutral values. Similarly, voice characteristics could be altered or various "accents" could be achieved. Implementation of speech rate variation in a realistic manner will require us to adjust the timing of formant trajectories. Different phases of vowel nuclei (onglide, steady state, offglide) are known to stretch and compact differently. A formant-based representation will enable us to identify and simulate these phases accurately and consistently.

There are also major problems with the formant representation. Formants can merge, split, cross, emerge and die, and available analysis tools are insufficiently reliable and consistent in their treatment of such situations. LPC based methods frequently produce parameter discontinuities requiring a hand tuning phase. We have developed a new tool based on arc-length minimization (ALM) which has given satisfactory results. This method is described in a separate ICLSP98 paper by the first author. The key features of this method are automatic operation, smooth and accurate formant parameters, and simultaneous pitch epoch marking.

Crossfade between concatenation units is confined to regions where the formants are stable. In other regions where the formant model fails, the requirement to exactly analyze the formants is relaxed, and filter parameters move to neutral and wide bandwidth values.

Since many of the units are redundant in various ways, we are in a position to reduce our overall database size by finding those redundancies and developing schemes to take advantage of these. Some standard methods include vector quantization and stylization of formant parameters, but we are also using methods that are inherent in this particular speech model. One approach relies on the idea that the voice source is devoid of most vowel difference information, and hence the voice source which is also generated by concatenation need not be represented for every vowel. We did tests to verify that the reduced voice source data was still adequate to generate all vowels. Additional experiments suggest that the filter parameter data can also be reduced. Although English has many vowels, the initial part of many vowels are similar to those of other vowels. A method was devised to take advantage of this by optimizing the cross-fading function, and hence allowing a reduction in initial half-syllables.

4. UNIT SELECTION

Unit selection involves determination of appropriate source and filter speech units for a sequence of phonemes. Selection is carried out on a syllable-by-syllable basis.

The speech unit database will be organized so as to expedite the search for the most specifically applicable unit for any given context. Source units, whose time-domain data requires considerable storage, will be more generalized than filter units; thus, /piht/, /peht/, /paht/ can be generated with the same pV- and -Vt source units, but require vowel-specific filter units. For either type, it is forseen that some contexts may require more specific treatment. To cover such cases an exception library will override the usual mechanism of unit selection, providing data for complete syllables and high frequency words. The size of this library can be varied for different applications to obtain an optimal balance of size and speech quality.

5. UNIT CONCATENATION

The construction of syllables requires concatenation and crossfading of two source demisyllables and two filter demisyllables. At the source level, crossfading can be realized over the course of most of the vowel, and we feel that crossfading over a large number of pulses is essential to prevent discontinuities. Filter units are crossfaded in the steady state portion of the vowel, where the formant values of the two can be expected to be close. Concatenation at syllable edges is similarly arranged around steady state portions of consonants; for voiceless consonants, simple overlap-add crossfading can be used.

Key phonetic events, such as voicing onset/offset, velar opening/closing, and articulator opening/closing, are aligned between the source and filter. The relative durations of the segments defined by these events are based on those in the original data, adjusted according to speech rate and other prosodic factors.

A key advantage of the current system is that crossfading of formant trajectories retains the identity of each resonance. Whereas models which crossfade in the time domain (such as TD-PSOLA) or which crossfade entire spectral envelopes (such as some types of harmonic synthesis) tend to blur formants, the current approach more closely approximates the formant transitions of natural speech.

6. PROSODIC MODIFICATION

The pitch control is handled in the source part of the source-filter model. This can be handled by various methods as in a previous paper [8], but is essentially a concatenation of glottal pulses, which are overlapped. A particular pulse can be chosen for inherent period, which is close to the desired period, or can be chosen for its proximity in the original context, which matches the desired context for synthesis. For example, to synthesize the first part of voicing in "pot", the glottal pulses are typically derived from the first part of a /peh/ demisyllable. The pitch epochs of the synthesized glottal pulse sequence are time-aligned to produce the desired pitch pattern.

The duration and timing control are again handled in the source. The preferred way is to overlap the two half-syllables to such an extent that the entire syllable has the desired duration. In the case of boundary units, we again adjust the overlap of crossfade between pieces, to produce the proper timing of events. In some cases a "self-cross-fade" is used. For example, if it is necessary to modify the duration of /l/ in the demisyllable /bla/, the demisyllable can be split into two part, /bl/ and /la/, and a cross-fade implemented during the /l/ to adjust its length.

7. RESULTS

In developing this approach we had three objectives. First, it should avoid the types of acoustic distortion and glitches associated with many synthesis methods. Second, it should be implementable as a relatively compact and computationally efficient system. Finally, it should produce speech which is natural and human sounding.

Given optimally analyzed speech units, the current method is almost completely free of acoustic distortion. The greater stability of formants eliminates the spectral warble associated with jumps of filter coefficients in LPC models, while avoiding the blurring of vocal tract resonances associated with some TD-PSOLA and harmonic-stochastic systems. The scratchiness of pure TD-PSOLA systems, attributed to phase mismatches, is also absent, and naturalness is retained even with extreme pitch distortion; we credit this to the lack of major resonances in the glottal source, and to improved pitch epoch marking. Noise components of the speech signal, such as in stop bursts and fricatives, are reproduced naturally, and the use of boundary units ensures that important transitional cues which occur at segment junctures are preserved.

Efforts to constrain the storage requirements of the system have also been successful. In initial investigations we have found that formant trajectories can be superimposed on voice source units originally derived from other vowel contexts without degradation of the resulting signal, and this ability to generalize drastically reduces the amount of time-domain data which must be incorporated into the system. Further, stylization of formant trajectories allows filter units to be compressed considerably without perceptual degradation.

Initial versions of the system have shown that a high degree of naturalness is attainable from the techniques described in this paper. Synthesized speech retains individual characteristics of the original speaker. This is true for both male and female voices, which is a striking improvement over traditional formant-based systems. We have also found the approach to work well in synthesis of segment types, such as nasals and nasalized vowels, which have been problematic for formant synthesis.

8. DISCUSSION

In the development of these ideas into a complete system, a number of points need to be addressed. The size of our unit database demands that automatic and reliable analysis methods be used. The number of possible crossfading combinations means that these methods must be highly consistent in their marking of pitch epochs, identification of formants, and marking of crossfade regions, in order to ensure smooth transitions. A second requirement is to evaluate the degree to which speech units can be generalized, and to determine optimum subsets of the whole database for small-footprint versions. A long-term objective will be to apply these methods to the synthesis of other languages.

REFERENCES

1. Allen J., Hunnicutt S., and Klatt D. "From text to speech: the MITalk system". MIT Press, Canbridge, Massachussetts, 1987.

2. Matsui K., Pearson S., Hata K, Kamai T. "Impproving naturalness in text-to-speech synthesis using natural glottal source". Proc. ICASSP, pp. 769-772, 1991.

3. Pearson S, Holm F., Hata K, "Combining concatenation and formant synthesis for improved intelligibility and naturalness in text-to-speech systems", International Journal of Speech Technology, Vol. 1, No. 1, pp. 103-107, 1997.

4. van Santen J., "Combinatorial issues in text-to-speech synthesis", In Proceedings Eurospeech 1997.

5. Moulines D. and Charpentier F. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". Speech Communication, vol. 9, no 5, pp. 453-467, 1990.

6. Serra X., Smith J., "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition". Computer Music Journal, Vol. 14, No. 4, 1990.

7. Imai S., Kitamura T. and Takeya H., "A direct approximation technique of log magnitude response for digital filters", IEEE Trans., vol. ASSP-25, No. 2, pp. 127-133, April 1977.

8. Pearson S.,Javkin H., Matsui K., Kamai T., "Text-to-speech Synthesis using a natural voice source", Proc. ICSLP 1990.