Published in *Proceedings IEEE 2002 Workshop On Speech Synthesis*, Santa Monica, CA.
For more recent work by Hertz and her collaborators, see http://www.novaspeech.com/publications.htm

INTEGRATION OF RULE-BASED FORMANT SYNTHESIS AND WAVEFORM CONCATENATION: A HYBRID APPROACH TO TEXT-TO-SPEECH SYNTHESIS

Susan R. Hertz

SpeechWorks International, Inc.
127 W. State St., 2nd Floor, Ithaca, NY 14850
and
Department of Linguistics, Cornell University, Ithaca, NY 14853

ABSTRACT

This paper describes an approach to speech synthesis in which waveform fragments dynamically produced with a set of formant-based synthesis rules are concatenated with pre-stored natural speech waveform fragments to produce a synthetic utterance. While this hybrid approach was originally implemented as a tool for research into improved voice quality in formant-based synthesis, it has produced such good results that we now view it as a potentially viable and advantageous approach for a text-to-speech product. Possible advantages of the approach include smaller speech databases for waveform concatenation, enhancement of certain speech cues for sub-optimal listening environments, and improved and more efficient unit selection/production. In addition, the approach has already proven its utility as a tool for research and development in both concatenative and formant-based synthesis.

1. INTRODUCTION

Our work in hybrid synthesis began through research on voice quality cues in formant-based synthesis. Specifically, we were working on the female voice produced by the formant-based ETI-Eloquence text-to-speech rule system for US English [1]. The ETI-Eloquence system produces highly intelligible speech segments and natural-sounding prosody, but the overall voice quality does not generally sound human, particularly in the case of the female voice. It was difficult to determine through synthesis alone how different segment types contributed to the unnatural quality of the female voice, so we decided to explore this question by splicing together synthesized and natural speech segments for US English utterances.

As expected, most stressed syllable nuclei from ETI-Eloquence sounded unnatural in the hybrid utterances, but many other types of segments sounded surprisingly natural in all or many contexts, in fact often indistinguishable from their natural speech equivalents. Obstruents worked best, but even many sonorants could be used in a wide range of contexts with no ill effects. Often as many as two-thirds of the phones within a sentence could be synthesized with little or no degradation in speech quality.

The following sections discuss the specific kinds of hybrid experiments we conducted, the preliminary results we obtained, and the possible advantages of the approach.

2. HYBRID EXPERIMENTS

We performed two types of hybrid experiments. In one, we constructed strategically selected sentences by hand using different combinations of synthesized and natural speech segments. In the other, we implemented a text-to-speech prototype that automatically produces speech based on specified combinations of synthetic and natural speech segment types.

2.1 Hand Construction of Utterances

We hand-constructed about thirty short- to medium-length utterances (approximately 300 words total) by concatenating natural and synthetic speech waveform fragments for adult male and female voices using the WSOLA join technique [2] employed in our concatenative waveform-based Speechify 2.0 text-to-speech system. Some of the hybrid utterances we constructed were ones for which there were original recordings in the database, so we could compare the hybrid utterances with completely natural speech. In other cases, we worked with utterances produced by the Speechify 2.0 text-to-speech system, which generates speech by concatenating half-phone-sized waveform fragments extracted from a natural speech database. All the Speechify utterances we selected sounded highly natural.

The natural waveform fragments used for hybrid utterance construction were all extracted from the 8 kHz versions of the speech databases for the Mara and Rick voices of Speechify 2.0. The synthesized waveform segments were generated at an 8 kHz sampling rate with the Reed and Shelley voices produced by the ETI-Eloquence 6.0 rules for US English. For each utterance, these rules produce parameter values for a formant synthesizer, which generates a synthetic waveform from the parameter values. The synthesizer is similar to the KLSYN-88 synthesizer described in [3], and uses a glottal source model much like KLGLOTT-88 [3], a cascade vocal tract model, and a parallel fricative model.

While the ETI-Eloquence rules are based on Hertz's phone-and-transition model [4], in which formant transitions are independent units, for purposes of our hybrid experiments we treated transitions as parts of phones, employing the same assumptions that underlie the segmentation strategy used for the Speechify databases. For example, voiced transitions between obstruents and sonorants were considered part of the sonorants; aspirated transitions out of stops were considered part of the stops; and transitions between vowels were divided evenly between the two adjacent vowels.

Using custom scripts, we constructed hybrid utterances that contained different combinations of natural and synthetic speech. For example, we generated different renditions of an utterance with only stressed vowels synthesized, with only unstressed vowels synthesized, with all sonorants synthesized, with only obstruents synthesized, etc. We also listened to a number of sentences in which a synthetic version of each phone was spliced in one by one, so that we could focus our attention on individual segments. In these cases, we ensured that the phones were correctly segmented in the natural speech databases for Mara and Rick.

After constructing hybrid utterances, we elicited informal judgments from about ten people, including both speech researchers and more listeners. Some people listened over headphones, others over speakers. Some listeners were presented with hybrid and/or natural utterances played back to back and asked which version they preferred; some were asked to point out anything that sounded odd about an utterance; and others were asked specifically which part of the utterance was synthetic.

Despite the informal and exploratory nature of the work to date, we are extremely encouraged by the fact that listeners often considered hybrid utterances containing a high percentage of synthetic segments to be equal in quality to the corresponding fully concatenative versions or natural speech recordings. Listeners, even those very familiar with ETI-Eloquence, also could not reliably determine which segments were synthetic, with the exception of relatively long syllable nuclei, which often sounded unnatural, especially in the female voice. The following examples show mixes of synthetic (bold underlined) and natural speech fragments for four highly natural-sounding hybrid utterances.

Mara + **Shelley**:

1. I**t** **w**a**s** o**bv**i**ous** **wh**y **the** **ch**ee**tah** a**te** **s**o **m**u**ch** **f**oo**d.**

2. La**st** year R**eb**e**cca** **t**oo**k** **th**e **sp**a o**ver** **and** **m**a**de** **it** **pr**o**f**i**table.**

Rick + **Reed**:

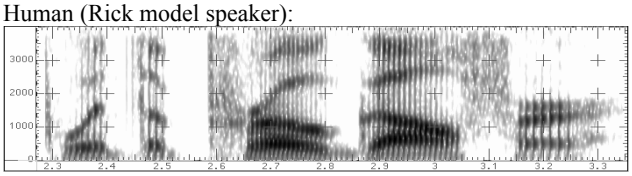
3. **Th**e e**xpert** **sk**ier a**gr**ee**d** **to** **t**ee u**p** **with** **the** **pr**o **g**ol**fer.**

4. **A** **Nebr**a**sk**a Au**gust** **is** **a** **g**oo**d** **t**i**me** **for** **picking** **b**err**ies.**

Figure 1: Sample hybrid sentences (synthetic segments bold underlined)

The following figure shows spectrograms of original human and hybrid renditions of the last four words of sentence 3 in Figure 1 above. Note, however, that spectral and durational differences do not necessarily correlate with perceived differences.

Human (Rick model speaker):



Hybrid (Rick + Reed):

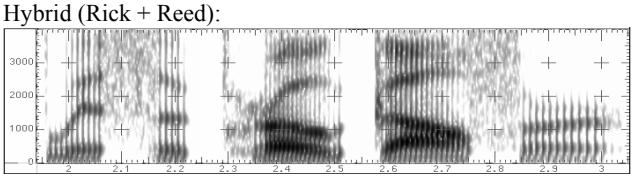


Figure 2: Human and hybrid renditions of *with the pro golfer*

Perhaps the most compelling result of our hybrid experiments was that synthetic renditions of non-nasal obstruents (i.e., voiced and voiceless stops and fricatives) generally sounded natural. Waveform joins were usually imperceptible at obstruent boundaries, and synthetic obstruents often sounded equally natural to their natural-speech counterparts, and sometimes even better. For example, in certain isolated words we presented, listeners preferred the synthetic renditions of [s], [t], and [d], presumably because the ETI-Eloquence rules lower the

frequency of the fricative noise for these segments when using an 8 kHz sampling rate so that the noise will be audible. This preliminary result suggests that one of the advantages of hybrid synthesis might be to enhance certain consonant cues for telephony applications or noisy environments.

Further, the relatively high frequency of non-nasal obstruents in English suggests that we can reduce the size of our speech databases by eliminating these segments and synthesizing them dynamically. In an analysis of the phone output produced by the ETI-Eloquence/Speechify front end for 160,225 randomly-selected English sentences comprising 11,967,343 total phones, non-nasal obstruents accounted for about 37% of the phones and 32% of their total duration, though the savings in any given database would depend on details of the utterances in that database. The formant-based rules themselves would add little memory: the entire speech generation component of ETI-Eloquence is less than 200K in size for all voices and all segment types, and further space reductions are possible.

The results concerning nasal obstruents and other types of consonants were less consistent than those for non-nasal obstruents, and more research is needed to determine the precise conditions under which they will work. We were surprised, however, by how many of the synthetic sonorant consonants we tried sounded natural.

Synthetic nasals sounded best in syllable onsets, and rarely worked well in syllable codas, especially when the syllable carried a pitch accent. This result is not surprising in light of the fact that pitch accents in English are generally realized over the entire vowel-nasal sequence, and pitch discontinuities may occur when a nasal from one pitch context is combined with a vowel from another. It might be possible to modify the nasals in these cases in accordance with the fundamental frequency and other properties of the preceding vowel to achieve better results, as we did for certain segments in other contexts (see below), but no attempt was made to do so in our preliminary experiments.

Like nasals, the sonorants [l] and [r] worked best in syllable onsets, which again is not surprising, since vowels plus tautosyllabic [l] and [r] form single durational, amplitude, and pitch units in English, as discussed in [5]. Also, vowel-liquid sequences in syllable nuclei are hard to segment reliably, so it is best not to break them up.

We examined selected segments that sounded unnatural in the hybrid utterances, and found that we could improve the naturalness by modifying certain synthesizer parameters. For example, some nasals in syllable codas sounded too nasal, and we successfully reduced the nasality by narrowing the second formant bandwidth. Such modifications not only improved the consonants in the hybrid utterances, but also in the fully synthetic ones, where we had not noticed the problems due to other, more salient, voice quality problems. In general, we found the hybrid approach to be a useful tool for identifying spectral and durational problems with segments in our formant-based rule system. Conversely, as discussed in section 2.2, we also found the approach to be useful in isolating problems in our waveform-based concatenative system.

One of the most surprising results of our work was the fact that so many synthetic syllable nuclei sounded natural within the hybrid utterances or produced minute quality degradations detectable only through back-to-back listening with natural speech. Those that worked consistently best were reduced vowels or reduced vowels plus sonorant sequences, particularly when the nucleus occurred between non-sonorants, where join problems are minimized and spectral discontinuities are less likely to occur. The best results were achieved when we adjusted the spectral tilt and aspiration amplitude of the synthetic nuclei to increase breathiness [3]. Occasionally, we needed to adjust the fundamental frequency as well. The ability to modify fundamental frequency appropriately for the context without degrading voice quality could be advantageous in a waveform concatenation system, since modifications made directly to the waveforms themselves tend to degrade the speech quality.

Collectively, the results suggest that the potential exists to synthesize dynamically at execution time a relatively high percentage of segments in most utterances. Further, since the hybrid utterances still sounded like Rick or Mara, the results suggest that the perception of voice quality stems very much from stressed syllable nuclei, and that listeners are less sensitive to other parts of the utterance. While we have not explicitly returned to the formant-based voice quality work that motivated the hybrid experiments to begin with, the results of our hybrid experiments point clearly to the areas we should address.

2.2 A Prototype Text-to-Speech System

Encouraged by the results of the hand-constructed utterances, we developed an experimental hybrid text-to-speech system that combines the 8 kHz ETI-Eloquence 6.0 Shelley and Speechify 2.0 Mara voices. With this system, we can specify in the textual input precisely which phones to synthesize by rule and which to extract from the natural speech database. Textual annotations can also be used to instruct the system to synthesize all segments of a particular type, to reduce certain amplitudes

by a specified number of dB, and to make other types of adjustments of the sorts suggested by the hand-construction experiments. The prototype runs under Windows 2000 on PCs and under Solaris 2.7 on Sun SPARCstations.

The author listened to several hundred sentences produced by the hybrid prototype back to back with the output of Speechify 2.0, adjusting synthesizer parameters as appropriate. Like the hand construction experiments, the hybrid prototype yielded promising results. However, it was more difficult to draw conclusions due to confounding factors such as non-optimal selection of natural units in certain cases or segmentation errors in the Mara database. For example, some stops selected by Speechify were not sufficiently aspirated for the context, in which case the synthetic versions generally sounded better. In another case, the end of a selected vowel was labelled prematurely in the database, and the incorrect duration of the vowel made the following synthetic voiceless stop sound voiced. Segmentation problems of this sort were not always noticeable in the fully concatenative versions, particularly when the segments in question were selected as a single chunk from the same utterance and not individually joined. In general, the hybrid synthesis proved useful for uncovering a number of database alignment problems.

We did not evaluate the impact on performance of the hybrid approach, since our prototype was not optimized for maximal run-time efficiency. However, in a product implementation, the time taken by the formant synthesis might be more than offset by the unit selection optimizations that the approach makes possible. For example, since the network of candidate units that Speechify produces for the target phones could be collapsed at synthesized insertion points, a significant reduction in search complexity could be achieved.

Besides its potential in text-to-speech products, the work with both the hand construction of hybrid utterances and the hybrid prototype has made clear that the hybrid approach is useful as a tool in its own right for research into concatenative waveform-based text-to-speech systems like Speechify. We might use it, for example, to evaluate in a systematic fashion where waveform joins are perceptible, to help detect database alignment problems, and to help identify the source of poor-sounding units. When a stop sounds unnatural, for instance, we could test hypotheses about the source of the problem by systematically replacing the stop with a synthetic one that is longer, shorter, more aspirated, has a stronger burst, etc., using much the same kind of experimental procedures employed in the development of the ETI-Eloquence rules. In general, the hybrid work as well as independent research not reported on here has shown that the perceptual results of experiments using formant synthesis often carry over directly to natural speech.

3. CONCLUSION

This paper has presented a hybrid approach to speech synthesis that combines formant-based synthesis by rule with natural speech waveform concatenation, and has suggested ways in which a hybrid-based text-to-speech system might capitalize on the strengths of each approach. Specifically, such a system would have more natural-sounding voice quality than current systems based entirely on rules, while other aspects of the speech could be improved through the ability to carefully control parameters in synthesized portions. Further, since a large percentage of segments could be generated on the fly, with a small set of rules, the overall size of the system could be reduced. The hybrid approach also has shown potential as a tool for evaluating and researching problems in both formant-based and waveform-based speech synthesis. Through our preliminary hybrid experiments, we have already gained new insights into the perception of voice quality and speech more generally, and we see much promise in a synthesis of the two technologies.

4. ACKNOWLEDGEMENTS

The author thanks her research assistants Bonnie Puckett and Steve Baker for their help with the hybrid hand construction experiments, her colleague Andy Wyatt for implementing the hybrid prototype, and her colleagues Rebecca Younes, Andy Wyatt, and Drew Lowry for helpful comments on the paper.

5. REFERENCES

[1] S.R. Hertz, R.J. Younes, and N. Zinovieva, "Language-Universal and Language-Specific Components in the Multi-Language ETI-Eloquence Text-to-Speech System," *Proceedings of the 14th International Congress of Phonetic Sciences,* vol. 3, pp. 2283-2286, August 1999.

[2] W. Verhelst, "Overlap-Add Methods for Time-Scaling of Speech," *SpeechCommunication*, vol. 30, pp. 207-221, 2000.

[3] D.H. Klatt and L.C. Klatt, "Analysis, Synthesis and Perception of Voice Quality Variations Among Female and Male Talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820-857, 1990.

[4] S.R. Hertz , "Streams, Phones, and Transitions: Toward a Phonological and Phonetic Model of Formant Timing," *Journal of Phonetics*, vol. 19, *Special Issue on Speech Synthesis and Phonetics*, (R. Carlson, Ed.), 1991.

[5] S.R. Hertz and M.K. Huffman, "A Nucleus-Based Timing Model Applied to Multi-Dialect Speech Synthesis by Rule,"

Proceedings 1992 International Conference on Spoken Language Processing, vol. 2, pp. 1171-1174, October 1992.