THE AT&T NEXT-GEN TTS SYSTEM

http://www.research.att.com/projects/tts

M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal
AT&T Labs – Research
Shannon Labs, 180 Park Ave.,
Florham Park, NJ 07932-0971
U.S.A.

ABSTRACT

The new AT&T Text-To-Speech (TTS) system for general U.S. English text is based on best-choice components of the AT&T Flextalk TTS, the Festival System from the University of Edinburgh, and ATR's CHATR system. From Flextalk, it employs text normalization, letter-to-sound, and prosody generation. Festival provides a flexible and modular architecture for easy experimentation and competitive evaluation of different algorithms or modules. In addition, we adopted CHATR's unit selection algorithms and modified them in an attempt to guarantee high intelligibility under all circumstances. Finally, we have added our own Harmonic plus Noise Model (HNM) back-end for synthesizing the output speech. Most decisions made during the research and development phase of this system were based on formal subjective evaluations. We feel that the new system goes a long way toward delivering on the long-standing promise of truly natural-sounding, as well as highly intelligible, synthesis.

1. INTRODUCTION

Text-to-Speech (TTS) systems are entering the mainstream of advanced telecommunications applications. For this, they have to deliver highly intelligible output for general text, while also sounding natural. In a Nov. 1998 TTS comparison test conducted by ESCA/COCOSDA, a total of 17 systems competed in the English language (13 US and 4 UK English systems representing female and male voices; among these systems were Microsoft's Whistler, British Telecom's Laureate, Lucent Bell Labs' TTS, and our Next-Gen TTS). Figure 1 depicts results of this comparison test in three important categories without explicit identification of the systems (as required by the competition rules). Figure 1(a) reports on "overall voice quality" using a rating scale from 1 to 15 for each of the systems (labelled 1 to 17). This category (out of the three) comes closest to our notion of "naturalness". In Fig. 1(b), intelligibility results are expressed in terms of percent correctly transcribed words in sentences that were designed to minimize contextual information. Finally, Fig. 1(c) shows listeners' overall impression, again on a scale from 1-15. First, note the wide variability in overall voice quality and the comparatively smaller variability in intelligibility scores. This implies that the intelligibility problem in TTS might be close to solved while naturalness still has plenty of room for improvement. Second, and more importantly, note that several systems showed respectable results in either voice quality or in intelligibility (percent words correct), but not in both. This

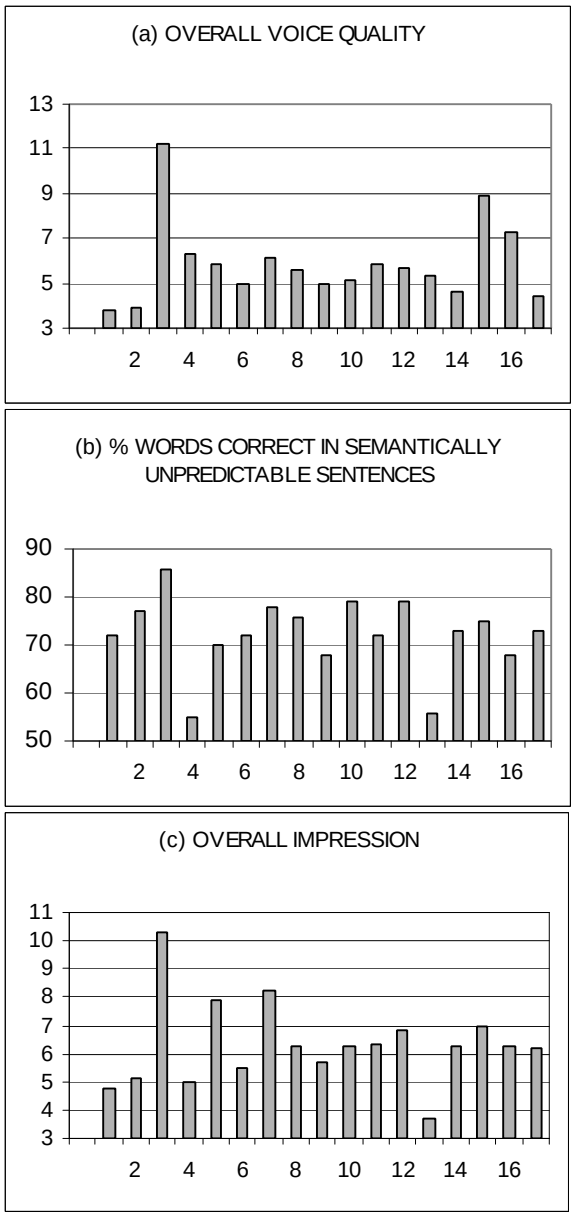

Fig. 1: Results from a recent TTS comparison test. (a) overall voice quality, (b) % of words correctly recognized from semantically unpredictable sentences, (c) overall impression.

finding reflects the relative strengths and weaknesses of individual systems. Finally, one system excelled in both categories, which led to its top rating in overall impression. We believe that such a quantum leap in quality is required in order to introduce TTS into a large variety of telecommunication services offerings.

Although test procedures were not totally satisfactory for many reasons, evaluations like the ESCA/COCOSDA tests will continue to be necessary in order to measure further improvements. The importance of thorough evaluations is an important point we stress in this paper. We also report on our efforts toward achieving, for general input text, higher naturalness in synthetic speech while, at the same time, maintaining high intelligibility.

The paper is organized as follows. Section 2 describes the architecture chosen for the AT&T Next-Gen TTS system. Section 3 introduces the notion of robust unit selection that insures previously unseen levels of naturalness while maintaining a high level of intelligibility. Section 4 summarizes our Harmonic-plus-Noise Model (HNM) synthesis back-end. Finally, section 5 stresses the importance of ongoing evaluation in the research and development of our system. We end the paper in section 6 with an outlook on further research.

2. SYSTEM ARCHITECTURE

Figure 2 shows a block diagram of the current system. AT&T's Next-Gen TTS is implemented within the Festival framework (CSTR, Univ. Edinburgh, Scotland). Text normalization, linguistic processing such as syntactic analysis, word pronunciation, prosodic prediction (phrasing and accentuation), and prosody generation (translation between a symbolic representation to numerical values of fundamental frequency F_0, duration, and amplitude) is done by a Flextalk object that borrows heavily from AT&T Bell Labs' previous TTS system, Flextalk. From ATR's CHATR system, we adopt the online unit selection (with modifications). Finally, speech is synthesized using one of many possible back-ends, including our own Harmonics plus Noise Model (HNM) synthesizer.

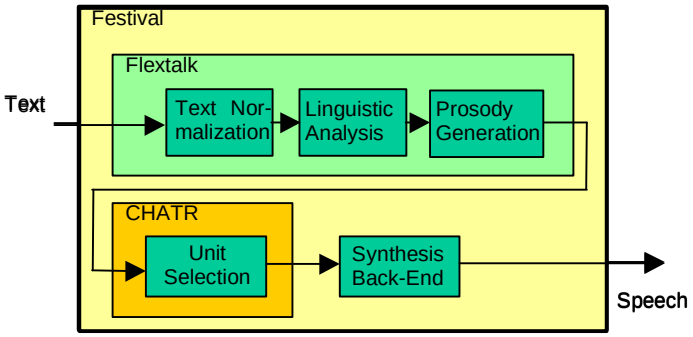

Fig. 2: System architecture of the AT&T Next-Gen TTS.

Note that the Festival architecture allows a remarkable degree of flexibility. Such flexibility helps considerably in producing comparative data. For example, we are able to test several synthesis back-ends with identical input. As another example, we can compare easily corresponding Flextalk and Festival modules while keeping all other components the same.

3. ROBUST UNIT SELECTION

The goal of unit selection is to provide a mechanism whereby segments of prerecorded speech are selected for synthesis. These segments are provided from a database. Many existing speech synthesis systems have a form of (online) unit selection that is very simple: there is only one prototypical unit for each kind of unit. In such systems, the real unit selection actually is done off-line, either by experts using knowledge-based, informed choices, or by clustering methods using formant-based or other distance measures for identifying a single unit that is the most compatible with potential neighboring units. Some more advanced systems have several units of each kind in their inventories that can be chosen for particular situations where wide variations (due to coarticulation) in the realizations of a phoneme are possible. Such systems still generally offer only limited online choices.

Why would one wish to complicate the process of choosing units for synthesis and additionally increase the size of a database? The answer lies in the belief that a system that provides only one possible unit for synthesis at any given point is detectably deficient. Such a belief is borne out by listening tests.

Given that speech synthesis may be improved by increasing the number of units in a voice database above the number required for phonetic accuracy it would seem appropriate to go further than previous systems and automate the process of providing large voice databases. This is what is done in the CHATR speech synthesis system.

So far we have not mentioned units, just segments of prerecorded speech. Diphone synthesis has been popular for a number of years, due to the high intelligibility that such systems provide. They have the ability to preserve some of the coarticulation effects that are present at phoneme boundaries. However, such systems are handicapped by having a large number of distinct units (in the order of 1000-3000 diphones, depending on language and phone-set chosen), for which it is not easy to create sufficiently large databases that capture all relevant co-articulatory effects. Statistics become even worse if we demand that all relevant prosodic variations are covered [1].

Unit selection synthesis, as used in the CHATR system, requires a set of speech units that can be classified into a *small* number of categories such that sufficient examples of each unit are available to make statistical selection viable. Hence, the original

CHATR system uses phonemes as units. To avoid problems of concatenation at phoneme boundaries a flexible join technique is employed that allows moving unit boundaries.

In order to arrive at a *robust* paradigm (i.e., that results in consistently high synthesis quality), we have modified the CHATR unit selection system. We have chosen to use half phones as the basic units of synthesis in a way that allows both diphone and phone-based synthesis, and mixtures thereof. Figure 3 illustrates the half-phone-based unit selection paradigm (note the subscripts "l" for the left half of a phone, "r" for a right half of a phone). This, naturally, results in greater complexity, but in compensation assures a synthesis intelligibility that is comparable to that of diphone synthesis while significantly increasing naturalness. Clearly, a high-quality database with accurate, multi-level labels is a crucial element of the system. More details can be found in [2].

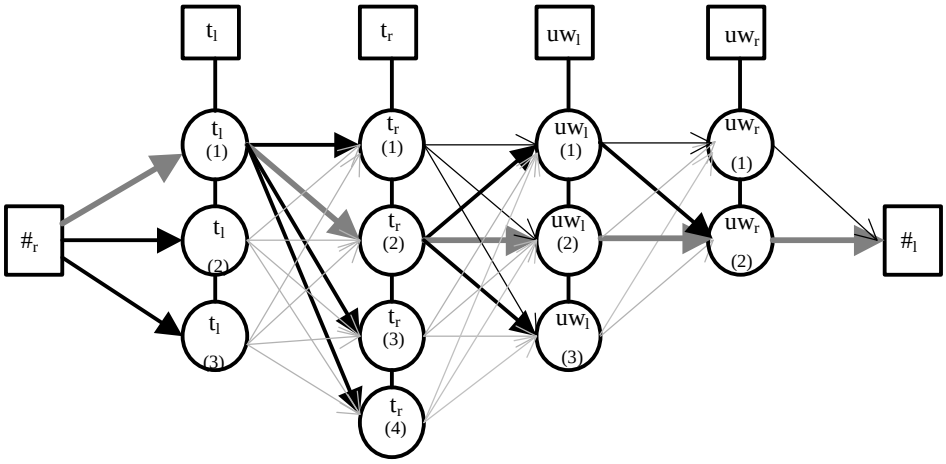

Fig. 3: Viterbi search based on an inventory of multiple instances of each half-phone needed for synthesizing silence -/t/-/uw/-silence (the word "two").

4. HNM SYNTHESIS BACKEND

The speech representation of choice for our female voice is the Harmonic plus Noise Model (HNM) [3]. In HNM the speech spectrum is divided into two bands: a low band, which is represented by harmonically related sinusoids with slowly varying amplitudes and frequencies, and a high band that is instantiated by a time-varying AR model that is excited by Gaussian noise.

HNM analysis consists of 3 steps. First, fundamental frequency F_0 and maximum voiced frequency (that determines the number of harmonics used) are set using a time-domain approach [4]. Then, harmonic amplitudes and phases are estimated by minimizing a weighted time-domain least-squares criterion. Finally, the AR filter for the high band is estimated by the autocorrelation approach. The analysis windows are set at a pitch-synchronous rate during voiced portions of speech and at a fixed rate during unvoiced portions. Note that HNM does not use predetermined pitch markers but estimates the length of local pitch epochs internally. Windows are two pitch epochs long.

For HNM synthesis, inter-unit phase mismatches are eliminated using the center-of-gravity approach [5]. Prosody (fundamental frequency F_0, duration, and amplitude) may be altered as desired (we do only limited modifications; see section 5). Around unit concatenation points, we smooth the HNM parameters in order to minimize residual discontinuities (after unit selection) by employing a simple linear interpolation over a small number of frames. The actual synthesis is done following the overlap-and-add paradigm. For each frame, the noise part is high-pass filtered according to the maximum voiced frequency found during analysis (that is zero for unvoiced speech). Also, the noise part is modulated by a parametric triangular envelope synchronized in time with the pitch period. Details can be found in [6].

5. EVALUATION

Formal listening tests were conducted throughout the research and development phase of Next-Gen TTS. We believe that selecting the voice for rendering the many hours of inventory speech was the most critical decision [7]. We also have identified acoustic correlates of listener ratings relevant to speaker selection [8]. Among several other crossroads decisions that were based on formal evaluations are the following: which back-end to use [9], which units (phones or diphones) lead to higher quality, and aspects of database pruning, post-lexical processing, etc. Details on these and other evaluations are given in [10] and [11].

Some important conclusions drawn from these tests are:

- A suitable speaker can have a very significant effect on voice ratings (of up to 0.3 points on a 5-point MOS scale in a comparison of several top-quality professional voices).
- In our current system, best performance is achieved with *no* prosody modifications during synthesis (other than some smoothing at unit boundaries) and relying on the prosodic coverage of the speech inventory instead. This led to a ~0.3 MOS point improvement in perceived voice quality (using HNM), as illustrated in Fig. 4. From such results it seems clear that, in

order to take advantage of modifications in F_0 to reduce the necessary speech inventory, we would need to know more about interactions between speech spectral envelope and F_0. Note, however, that predicted ("system") prosody is still used as a key to find the optimal sequence of units from the online inventory, but is not imposed on the selected units. Note also that the prosody modification experiment used the system's predicted prosody, which might indicate certain deficiencies of the prosody generation model. More on this and other experiments can be found in [11] and in [8].

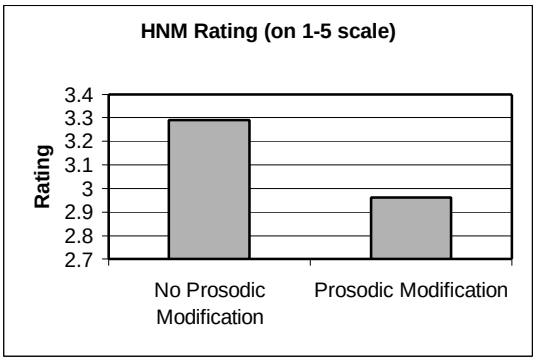

Fig. 4: Effect of prosody modifications during our version of unit-selection synthesis.

- Post-lexical processing (PLP), the expansion of acceptable pronunciations due to context, speaking style, dialect, or speaker idiosyncrasies, has a significant effect on ratings. Without it, a good sequence of units in the online inventory may not be selected at all because it does not precisely match the predicted transcription.
- It is not easy to "prune" an online database without sacrificing synthesis quality. Much more work is needed before we are able to say which portions of the database can be safely discarded. CHATR-style pruning of a total of 20% outlier units reduced synthesis quality by a small, but significant amount (~0.05 MOS points).

6. CONCLUSION

This paper introduced the AT&T Next-Gen TTS system. Key design goals were high voice quality and intelligibility. Also important were modularity, a modern implementation that adheres to the latest software engineering guidelines, and the fact that it is based on a generally available backbone architecture (Festival) that eases the learning curve of new team members.

In the future, more research will be done on speeding up the voice creation process with no degradation of synthesis quality, and on new voices in different speaking styles, accents, and languages.

7. REFERENCES

[1] J.P.H. van Santen (1997) "Combinatorial issues in text-to-speech synthesis." In: Proc. Eurospeech '97, pp. 2511-2514.

[2] A. Conkie (1999) "A robust unit selection system for speech synthesis." In: Proc. 137th meet. ASA/Forum Acusticum, Berlin, March 1999.

[3] Y. Stylianou, T. Dutoit, J. Schroeter (1997) "Diphone concatenation using a Harmonic plus Noise Model of speech." In: Eurospeech '97, pp. 613-616.

[4] Y. Stylianou (1996) "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech," IEEE Nordic Signal Processing Symposium, Helsinki, Finland, Sept. 1996.

[5] Y. Stylianou (1998) "Removing phase mismatches in concatenative speech synthesis." In: The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998, Paper H.2.

[6] Y. Stylianou (1998) "Concatenative speech synthesis using a Harmonic plus Noise Model." In: The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998, Paper H.1.

[7] A. Syrdal, A. Conkie, Y. Stylianou, J. Schroeter, L. Garrison, and D. Dutton (1997) "Voice selection for speech synthesis," J. Acoust. Soc. America, **102**, paper 4pSP5, p. 3191 (A).

[8] A. Syrdal, A. Conkie, and Y. Stylianou (1998). "Exploration of acoustic correlates in speaker selection for concatenative synthesis." In: Proceedings of ICSLP 98, Paper Number 882.

[9] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter (1998) "TD-PSOLA versus Harmonic plus Noise Model in diphone based speech synthesis," IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1998, Seattle, WA, pp. 273-276.

[10] A. Syrdal, G. Möhler, K. Dusterhoff, A. Conkie, and A. Black (1998) "Three methods of intonation modeling." In: The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998, Paper J.3 (R54).

[11] M. Beutnagel, A. Conkie, and A. Syrdal (1998). "Diphone synthesis using unit selection." In: The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998, Paper F.2 (R52).