RECENT ADVANCES IN MULTILINGUAL TEXT-TO-SPEECH SYNTHESIS

Bernd Möbius, Juergen Schroeter, Jan van Santen, Richard Sproat, Joseph Olive
AT&T Bell Laboratories, Murray Hill, NJ, USA

INTRODUCTION

In this paper we will discuss recent advances in multilingual text-to-speech (TTS) synthesis research at AT&T Bell Laboratories. The TTS system developed at AT&T Bell Laboratories generates synthetic speech by concatenating segments of natural speech. The architecture of the system is designed as a modular pipeline where each module handles one particular step in the process of converting text into speech. Besides conceptual and computational advantages, the modular structure has been instrumental in our effort toward a TTS system for multiple languages. This system will ultimately consist of a single set of modules, and any language-specific information will be represented in tables. In describing the TTS system, we will concentrate on the multilingual aspect, with a bias toward the German language.

A MODULAR ARCHITECTURE

The architecture of the Bell Labs TTS system is entirely modular. This design has a number of advantages (see [18] for a more detailed discussion) for system development and testing, and research. First, although the division of the TTS conversion problem into subproblems is always arbitrary to some extent, each module still corresponds to a well-defined subtask in TTS conversion. Second, from the system development point of view, members of a research team can work on different modules of the system, and an improved version of a given module can be integrated anytime, as long as the communication between the modules and the structure of the information to be passed along is defined. Third, it is possible to interrupt and (re-)initiate processing anywhere in the pipeline and assess TTS information at that point, or to insert tools or programs that modify TTS parameters.

The current English version of our TTS system consists of thirteen modules (Figure 1). Information flow is unidirectional, and each module adds information to the data stream. Inter-module communication is performed by way of a uniform set of data structures. The output of the unit concatenation module, however, is a stream of synthesis parameters; this information is finally used by the waveform synthesizer.

The modular architecture has been instrumental in our effort to develop TTS systems for languages other than English. Currently, work is under way on nine languages: Mandarin Chinese, Taiwanese, Japanese, Mexican Spanish, Russian, Romanian, Italian, French, and German.

In the initial stages of work on a new language much of the information needed to drive these modules is missing. Typically, we start with a phonetic representation of the phoneme system of the language in

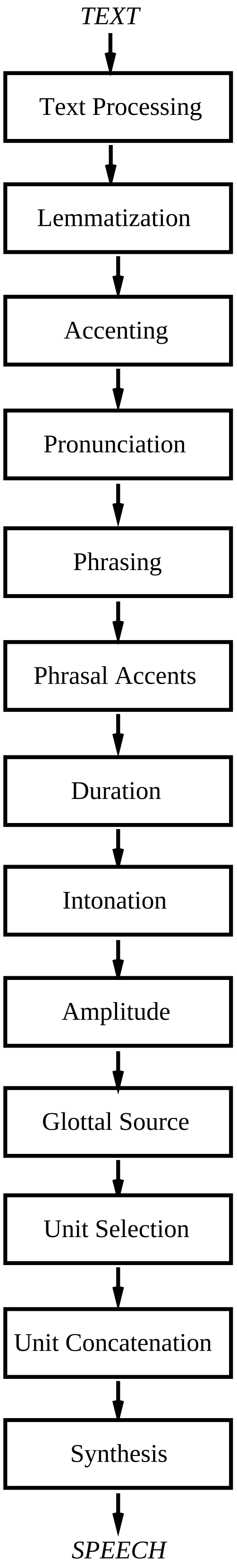

Figure 1. Modules of the English TTS System.

question and build the acoustic inventory, whereas text analysis and prosodic components would be worked on in a later stage. In this case, default versions of the modules enable the researcher to get a reasonable synthetic speech output for a given input string which can consist of phone symbols and control sequences for various prosodic parameters.

While some modules, such as unit selection, unit concatenation, and waveform synthesis, have already been largely table-driven for some time, we recently integrated language-independent text analysis, duration, and intonation components. Thus, the Bell Labs TTS system can now be characterized as consisting of one single set of modules, where any language-specific information will be represented in, and retrieved from, tables.

We will now turn to a discussion of components in the light of multilingual TTS.

TEXT ANALYSIS

The first stage of TTS conversion involves the transformation of the input text into a linguistic representation from which actual synthesis can proceed. This linguistic representation includes information about the pronunciation of words (including such ‘non-standard’ words as numerals, abbreviations, etc.), the relative prominence (accenting) of those words, and the division of the input sentences into (prosodic) phrases. Deriving each of these kinds of information presents interesting problems that can be solved using techniques from computational linguistics.

In the English TTS system, the information of how to pronounce a word resides in the pronunciation model itself. The model is basically a large set of word-specific pronunciation rules because it consists of a list of words and their pronunciations. An alternative approach [16] has been taken in the multilingual TTS system. Here lexical information is represented by (mostly morphological) annotations of the regular orthography. The generalized text analysis component computes linguistic analyses from text using a lexical toolkit that is based on state-of-the-art weighted finite-state transducer technology.

First, input text is converted into a finite-state acceptor which is then composed sequentially with a set of transducers that go from the surface representation to lexical analysis. Since this yields all possible lexical analyses for a given input, a language model helps find the presumably ‘correct’ or most appropriate analysis. The best path through the language model is then composed with a transducer that goes from lexical analysis to phonological representation and pronunciation (Figure 2).

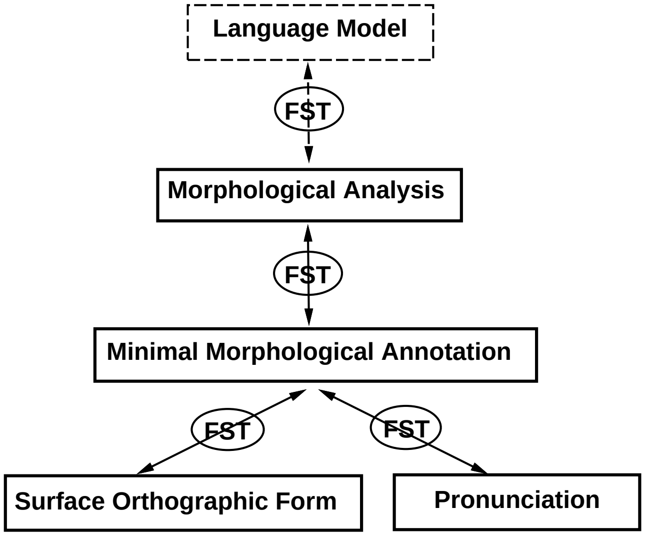

Figure 2. Generalized text analysis component for multilingual TTS based on finite-state transducer (FST) technology.

To illustrate the relevance of morphological information for pronunciation, let us consider two examples from German. One of the rules of German pronunciation is that an /s/ preceding a /p/ followed by a vowel or liquid is pronounced as [ʃ]. This rule only applies, however, if the /s/ is part of the same morpheme as the following /p/, and this in turn can cause a problem in compounds such as “Sicherheitspanne” (in principle potentially either “Sicherheit+s+Panne” or “Sicherheit+Spanne”). Since compounding is highly productive in German, most particular instances of compounds that one encounters in text will not be found in the dictionary. In a case like “Sicherheitspanne” morphological analysis tells us that the /s/ must be the ‘Fugen-/s/’ which regularly follows the noun-forming suffix “-heit”, and therefore should not be pronounced as [ʃ] but as [s].

In the second example, the input string “Sucht” yields two possible morphological analyses: s’uch{++}t{verb}{3per}{sg}{pres}{indi} with the pronunciation [z’u:xt] *vs.* s’ucht{noun}{femi}{sg}({nom}|{gen}|{dat}|{acc}) with the pronunciation [z’ʊxt]. Given equal probabilities for both alternatives, only information provided by a part-of-speech tagger or parser (e.g., [1]) can help disambiguate and determine the correct pronunciation.

The text analysis component also performs a tokenization of the input text into sentences and words. While some writing systems, e.g. Chinese, use a special symbol to mark the end of a sentence and nothing else, the situation is less fortunate in other writing systems. In German and many other European languages, a period is ambiguous in that it delimits sentences but also marks abbreviations. End-of-sentence detection, abbreviation, acronym and number expansion, word tokenization and other preprocessing problems are typically solved using a set of heuristics (e.g., [2] [10]).

Other linguistic information as derived by text processing includes information on parts of speech as well as on phrasing and accenting, and jointly forms the input to subsequent modules: *segmental duration*, *intonation*, *unit selection* and *concatenation*, and *synthesis*. In our multilingual systems, the new generalized text analysis component replaces all the modules up to *segmental duration*.

SEGMENTAL DURATION

The duration module assigns a duration to each phonetic segment. Given the string of segments to be synthesized, each segment is tagged with a feature vector containing information on a variety of factors, such as segment identity, syllable stress, accent status, segmental context, or position in the phrase. The module as such is language-independent, with all language-specific information being stored in tables. Table construction is performed in two phases: inferential-statistical analysis of the speech corpus, and parameter fitting.

In the case of the German TTS system, we first designed a factorial scheme, i.e. the set of factors and distinctions on these factors that are known or expected to have a significant impact on segmental durations. An important requirement was that the factors can be computed from text. We then applied a quantitative duration model that is implemented as a particular instantiation of a 'sums-of-products' model [12] whose parameters are fitted to a hand-segmented speech database [5]. This approach uses statistical techniques that are able to cope with the problem of confounding factors and factor levels, and with data sparsity.

During analysis, the segments are classified according to their position in the syllable (onset consonants, nuclei, coda consonants). Within these classes, subcategorizations were made in terms of phone types (e.g., voiceless stops, voiced fricatives, nasal consonants, etc.).

The data show rather homogeneous patterns in that speech sounds within a given phone class generally exhibit similar durational trends under the influence of the same combination of factors. Among the most important factors are: a) syllable stress (for nuclei, and to some extent for stops and fricatives in the onset); b) word class (for nuclei); c) presence of phrase and word boundaries (for coda consonants, and to some extent for nuclei). The analysis yields a comprehensive picture of durational characteristics of one particular speaker.

INTONATION

The intonation module computes a fundamental frequency contour (F_0) by adding three types of time-dependent curves: a phrase curve, which depends on the type of phrase, e.g., declarative vs. interrogative; accent curves, one for each accent group (accented syllable followed by zero or more non-accented syllables); and perturbation curves, which capture the effects of obstruents on pitch in the post-consonantal vowel.

This approach shares some concepts with the so-called superpositional intonation models that have been applied to a number of languages (e.g., [3] [6]). These models analyze the F_0 contour as a complex pattern that results from the superposition of several components, each of which has its own temporal domain. The weakness of the superpositional models is generally seen in their lack of precision as far as the alignment of the F_0 contour with the internal temporal structure of the accent group is concerned.

The key novelty in our approach, however, is that we model in detail how the accent curves depend on the composition and duration of the accent groups. This is important because listeners are sensitive to small changes in alignment of pitch peaks with syllables. Previous findings on segmental effects of timing and height of pitch contours are integrated in the new model [13]. Similar to duration module construction, modeling these dependencies involves fitting of parameters to a speech corpus.

ACOUSTIC INVENTORY

The majority of units in the acoustic inventory are diphones, i.e., units that contain the transition between two adjacent phonetic segments, starting in the steady-state phase of the first segment and ending in the stable region of the second segment. Units to be stored in the acoustic inventory are chosen based on various criteria that include spectral discrepancy and energy measures. Contextual or coarticulatory effects can require the storage and use of context-sensitive 'allophonic' units or even of triphones [7].

For example, the current acoustic inventory of the German TTS system consists of approximately 1250 units, including about 100 context-sensitive units. This inventory is sufficient to represent all phonotactically possible phone combinations for German. However, it will have to be augmented by units representing speech sounds that occur in common foreign words or names, e.g., the interdental fricatives and the /w/ glide for English, or nasalized vowels for French.

For acoustic inventory construction we use a new procedure [14] that performs an automated optimal element selection and cut point determination. The approach selects elements such that, for a given vowel, spectral discrepancies between elements for that vowel are jointly minimized, and the coverage of required elements is maximized. A toolset is provided that helps reduce the amount of manual labor involved in the selection of inventory elements.

Elements that have been selected for inclusion in the inventory are then extracted ('cut'), normalized in amplitude, indexed and stored in tables as acoustic inventory elements. The normalization done on a given element depends on the synthesis method used (see following section) and on the speech sounds involved in the element.

SELECTION, CONCATENATION, SYNTHESIS

The unit selection and concatenation modules select and connect the acoustic inventory elements. These modules retrieve the necessary units, assign new durations, pitch contours and amplitude profiles and pass parameter vectors on to the synthesis module which uses one of the synthesis methods described below to generate the output speech waveform.

The parametric waveform synthesis module provides flexible engines to assure the highest quality speech output for a given hardware platform and number of parallel channels running on that platform. Since usually more than 60% of the computational effort of the total TTS system is spent on waveform synthesis, hardware constraints can be met most easily by trading off quality *vs.* complexity in the algorithms used by the synthesizer.

Our TTS system uses vector-quantized LPC and a parametrized glottal waveform for synthesis. More recently, we have introduced a mixed formant/LPC representation and added waveform synthesis engines with varying degrees of complexity.

Each specific synthesis method requires its corresponding analysis scheme. In all cases, we use the pitch-synchronous analysis outlined in [19] with the option of using a mixed LPC/formant spectral representation [8]. The glottal waveform model used in the standard back end is that of Rosenberg [11]. Spectral tilt is implemented as a separate first-order FIR filter. Aspiration noise is added in the glottal open phase. Plosive transients are synthesized using the original LPC residual [9]. We also have the option to use the LPC residual throughout synthesis. For waveform synthesis schemes, we modified the analysis procedure to extract intervals of exactly two pitch periods (for voiced sounds; fixed 5 ms pseudo-periods for unvoiced sounds), and store the Hanning-windowed speech waveform. The system can, in principle, also accommodate articulatory parameters [4] [15] given that analysis and resynthesis yield sufficiently high quality.

SUMMARY

We presented recent advances and developments in our ongoing effort toward a TTS system with multilingual capability. The Bell Labs TTS system can also drive a ‘talking face’, a visual speech synthesis module that provides ‘lip-reading’ cues to enhance discrimination between confusable consonants such as nasals, or between labial and alveolar stops. In addition, a ‘talking head’ contributes a visual personality to an application such as a computer’s help system. In summary, we feel strongly that TTS systems have started to play an important role in everyday human-machine communications. In the future, TTS will sound increasingly ‘natural’ where desired, and will talk in several languages while conveying several speaker personalities in each language.

REFERENCES

[1] Church K. (1988): “A stochastic parts program and noun phrase parser for unrestricted text”. *Proc. 2nd Conf. on Applied Natural Language Processing*, 136-143

[2] Coker C., Church K., Liberman M. (1990): “Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis”. *Proc. ESCA Workshop on Speech Synthesis* (Autrans, France), 83-86

[3] Fujisaki H. (1983): “Dynamic characteristics of voice fundamental frequency in speech and singing”. In P.F. MacNeilage (ed.), *The production of speech* (Springer, Berlin), 39-55

[4] Gupta S., Schroeter J. (1993): “Pitch-synchronous frame-by-frame and segment-based articulatory analysis-by-synthesis”. *Journal of the Acoustical Society of America* **94**, 2517-2530

[5] Kiel Corpus (1994): The Kiel corpus of read speech, Vol. 1 (CD-ROM, Univ. Kiel)

[6] Möbius B. (1993): *Ein quantitatives Modell der deutschen Intonation – Analyse und Synthese von Grundfrequenzverläufen* (Niemeyer, Tübingen)

[7] Olive J.P. (1990): “A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds”. *Proc. ESCA Workshop on Speech Synthesis* (Autrans, France), 25-29

[8] Olive J. (1992): “Mixed spectral representation – Formants and linear predictive coding (LPC)”. *Journal of the Acoustical Society of America* **92**, 1837-1840

[9] Oliveira L.C. (1993): “Estimation of source parameters by frequency analysis”. *Proc. Eurospeech-93*, 99-102

[10] Pavlovcik P. (1991): FREND reference manual (Technical Report, AT&T Bell Laboratories)

[11] Rosenberg A.E. (1971): “Effect of glottal pulse shape on the quality of natural vowels”. *Journal of the Acoustical Society of America* **48** 583-590

[12] van Santen J.P.H. (1994): “Assignment of segmental duration in text-to-speech synthesis”. *Computer Speech and Language* **8**, 95-128

[13] van Santen J.P.H. (1996): “Segmental duration and speech timing”. In Y. Sagisaka, W.N. Campbell, N. Higuchi (eds.), *Computing Prosody* (Springer, New York)

[14] van Santen J.P.H., Möbius B., Tanenblatt M. (1994): New procedures for constructing acoustic inventories (Technical Report, AT&T Bell Laboratories)

[15] Schroeter J., Sondhi M.M. (1994):“Techniques for estimating vocal-tract shapes from the speech signal". *IEEE Trans. Speech and Audio Proc.* **2**(1) Part II, 133-150

[16] Sproat R. (1995): “A finite-state architecture for tokenization and grapheme-to-phoneme conversion for multilingual text analysis”. In *From text to tags: Issues in multilingual language analysis. Proc. ACL SIGDAT Workshop* (Dublin, Ireland), 65-72

[17] Sproat R., Olive J. (1995): “Text to speech synthesis”. *AT&T Technical Journal* **74**(2), 35-44

[18] Sproat R., Olive J. (1996): “A modular architecture for multi-lingual text-to-speech”. In J. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), *Progress in speech synthesis* (Springer, New York)

[19] Talkin D., Rowley J. (1990): “Pitch-synchronous analysis and synthesis for TTS systems”. *Proc. ESCA Workshop on Speech Synthesis* (Autrans, France), 55-58