ENGLISH TEXT-TO-SPEECH
AS A FUNCTION OF
CONCATENATING DIGITIZED SYLLABLES

Walter G. Rolandi
Horizon Research, Inc.

Program in Applied Linguistics
Boston University

ABSTRACT - This paper describes preliminary results obtained in an application of some more fundamental core technology. The core technology is the syllabic representation of English. An effort is underway to computationally determine (essentially) all of the syllables that collectively comprise the English language. While rules depicting syllabification in English have been described (Chomsky & Halle, 1968), an actual list of the language's constituent syllables does not appear to exist.

Identifying the syllables of English may have several implications for the speech science community. Some are indicated below. This paper discusses the potential for improvement in English text-to-speech applications. The initial results by no means imply revolutionary breakthroughs. On the other hand, initial results do suggest a substantial improvement over some existing text-to-speech methods.

INTRODUCTION

Most text-to-speech methods currently available can be grouped into three major categories. One employs digitizations of entire words which are relevant to some particular domain. For instance, text-to-speech systems for telephone banking transactions are now becoming common place. The system designers have exhaustively determined all of the words needed for all of the transactions supported by the system. Each word is individually digitized and sentences are constructed dynamically by concatenating digitizations of individual words. While obviously not "natural" human speech, these systems provide a relatively high degree of fidelity to human utterances. The major problem with a more general application of this method is that it would require digitizing all of the utterances of the application language. In English, this could require digitizing as many as 350,000 words.

On the opposite end of the "length of component" spectrum are systems which concatenate on the phoneme level. Inexpensive versions are commercially available that map English strings onto phonemic representations and subsequently construct vocalic representations by means of concatenating synthesized sounds which are associated with each individual phoneme (COVOX, 1988). While often impressive, this method will typically result in the now stereotypical "robot" speech. The speech is characterized by its artificiality which is due in no small part to the fact that all of the phonemes are synthesized by the machine. That is, the phoneme are not sampled human speech but rather are pre-stored units of synthetic data.

On the same level of granularity, more natural sounding speech can be obtained by using digitizations of human speech, broken down on the phonemic level. This was initially investigated but later abandoned. The problems were two-fold. First, phonemes that were digitized in one verbal environment entailed variations from the same phoneme digitized in another speech context. For example, the schwa sound in the first syllable of "arrive" differs slightly from that in the first syllable of "appeal" due to the phonetic influence of the "r" sound which it precedes. This contributes one abnormality to the speech so produced. Second, the issue of how best to obtain digitized samples arose. Simply recording a speaker enunciating "p" or "e" showed little promise. In addition to being, by definition, unnatural speech acts, (people do not typically make such sounds in isolation), this method produced abnormalities in between phonemes. It was difficult to determine where individual phonemes start and stop. This problem could be described as "determining the *essence* of a particular phoneme". That is, that part of a phoneme's signal which humans will predictably and collectively recognize as that phoneme, when present. Informal testing with co-workers suggested great variation. Extracting sample phonemes from digitized intact utterances was subsequently investigated. That is, having a speaker say an entire sentence, then with the use of a wave editor, cutting out instances of individual phonemes. This approach had its limitations as well. At least with the tools at hand, it was exceedingly difficult to accurately isolate usefully representative incidents of phonemes within a digitized utterance. The problem of where individual phonemes begin and end arises when cutting phonemes out of some longer utterance as well. Ironically, the fact that contextual variations were *not* evident in digitized isolated phonemes contributed to their abnormal features. The fact that they are evident in samples cut out of intact utterances, however introduces its own abnormality. Contextual features of phoneme variations sound odd when employed out of their proper context. Again, informal testing showed that there was great variation in the way listeners responded to combined phonemes. Their was no apparent consensus due to the introduction of artificial factors.

One way of overcoming this problem is to digitize all of the phoneme contexts and concatenate these as base units. This is apparently the method currently used with the Bell Laboratories Text-to-Speech (TTS) system (Hirschberg, J., 1990a). This approach, while obviously artificial, shows marked improvement over concatenation on the phoneme level, both synthesized and natural phonemes. The fact remains however, that more natural sounding speech is apparent when a human voice was used and samples are digitized on the *word* level. Because, as stated before, it would be prohibitive to try to digitize all of the words of a language, another, more atomic level of granularity was sought. While there may be many, every language can be characterized by a finite set of syllables. Speech acts on the syllabic level largely eliminate the issue of phonemic context because the syllable carries with it the phonetic environment of its phonemic constituents. The syllable became the focus of analysis.

METHOD

Before one can digitize the syllables of a language, those syllables must first be identified. While algorithms for determining the boundaries of English syllables are available, (Chomsky & Halle, 1968; Mackay, 1987; Akmajian, Adrian, Demers, & Harnish, 1984), there has apparently never been an attempt to exhaustively determine and catalog the set.

In order to pursue this preliminary goal, two basic processes are now being explored to determine the syllabic and phonemic content of a string. Each is embodied in an individual computer program. The first phase of processing uses a knowledge base of several hundred rules which map English strings into their phonemic equivalents. These rules were originally encoded by researchers at the Naval Research Laboratory and are now in the public domain (NRL Report 7948, 1976). English text is the *input* and phoneme strings are the *output*. This portion of the system is being continuously refined. Initially, its error rate was as high as 40%. Rule refinement has reduced this rate to about 15% so far.

The second phase transformation takes as *input* the *output* of the first program. The program examines the phonemic representation and divides it into syllables. It does so using rules which describe consonants that covary in English speech, the locations in words where they are permitted, and how they are divided along syllabic borders. When this work is complete, an (essentially) exhaustive list of syllables that define spoken English should be computable. Note that the term "essentially" is used in order to avoid the assertion that a definitive list will be computable. The set of syllables which will result will in large part be determined by the word list or corpora ultimately used as input data to the first program. An industry standard spelling dictionary is currently in use for testing and development purposes. The dictionary has about 24,000 entries.

Upon obtaining a substantial set of constituent syllables, they will be digitized and stored. A final program will be produced that will create the syllabic representation of English strings, look-up the stored constituent digitizations, and output an utterance by concatenating the stored digitizations.

INITIAL FINDINGS

Based on preliminary findings, it is important to point out that it is unlikely that this research will revolutionize text-to-speech methodology. All, or many of the peripheral problems in text-to-speech such as pronunciation speed, pitch, accent, and intonation are unaddressed by this approach. (see Hirschberg, J., 1990b; Olive, J. P., and Liberman, M. Y., 1985). The purpose of the research is to determine whether or not the approach supports sufficient improvement over existing methods to merit further investigation. Informal testing however has suggested that the number of listener errors in response to sample words created by concatenating digitized syllables is substantially less than when using samples on the phoneme level. The results much more closely approximate speech on the word level than on the phoneme level.

Much work remains to be done. One important task that needs to be completed is the determination of the number of syllables that reasonably constitute English. Initial finding report roughly 4000 one word syllables, most of which can be combined to form many other words. If, on the other hand, it is eventually discovered that more than 10,000 syllables are necessary to represent a complete range of English words, the project will have to be re-evaluated from a practical standpoint.

OTHER SPEECH SCIENCE APPLICATIONS

In addition to the academic value of cataloging the syllables of English, and in addition to the promising application in text-to-speech generation, this work may potentially impact other Speech Science areas and applications. Viewing speech acts as syllabic concatenations, one can pursue the analysis of speech acts *computationally* with reference to the auditory properties of the utterances involved. Syllabic representation of English will permit computational manipulation of English strings in terms of both the syllabic and phonemic constituents of the string. It will permit computers to respond to strings in ways similar to speakers and listeners: on the basis of their phonetic properties.

Some areas of potential impact under current exploration are:

> Intelligent word recognition and/or spelling correction based on phonological discriminations. Intelligent word processing, text retrieval, and text handling routines in data processing.
>
> Use as a tool in the analysis of human speech processing.
>
> Teaching machine applications where it is desirable to make reinforcement contingent on the phonetic approximation of correctly spelled words. Learning to spell and learning to write applications.
>
> Modeling language acquisition by computer as a function of syllable association.
>
> Speech recognition on the syllabic level of analysis.

REFERENCES

Akmajian, A., Adrian, A., Demers, R.A., and Harnish, R.M. (1984). *Linguistics: An Introduction to Language and Communication.* Cambridge, MA: MIT Press.

Chomsky, N., Halle, M. (1968). *The Sound Pattern of English.* New York: Harper & Row.

COVOX, Inc. (1988). *Speech Thing User Manual.* Eugene, Oregon.

Hirschberg, J., (1990a). Conversational communication at **AAAI 1990**, Boston.

Hirschberg, J., (1990b). "Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech". *Proceedings, Eighth National Conference on Artificial Intelligence.* Volume II, p.952. Cambridge, MA: MIT Press.

Mackay, I.R.A. (1987). *Phonetics: The Science of Speech Production.* Boston: Little, Brown and Company.

Naval Research Laboratory, (1976). "Automatic Translation of English Text to Phonetics by Means of Letter-to-Sound Rules". *NRL Report 7948.* National Technical Information Service: Document "AD/A021 929".

Olive, J. P., and Liberman, M. Y. (1985). "Text to Speech -- an overview", *Journal of the Acoustic Society of America, Suppl. 1, 78(Fall):s6.*

SPEECH WORKSTATION FOR ITALIAN TEXT TO SPEECH DEVELOPMENT

G. Abbattista, A. Riccio, S. Terribili

ALCATEL FACE STANDARD, Research Centre
Pomezia (Rome) - ITALY

ABSTRACT - The paper describes the implementation and use of a powerful Workstation suitable for the generation of the acoustic units database and the study and evaluation of the prosodic contours for a Text to Speech system for Italian language.

INTRODUCTION

The development of a text to speech system, based on the segment concatenation approach, requires for the acoustical part, to build up a database of speech elements; these elements have to be extracted from real speech, coded in a convenient technique, and stored together with some additional information necessary for the concatenation. Generally this process is highly time consuming, it requires an expert operator and possibly a good level of interaction to allow corrections, afterthoughts and similar needs; moreover the final quality of the complete system is strongly affected by the appropriateness of this part of the development.
It is obvious that to overcome these problems the optimal solution would be an integrated environment provided with all the necessary tools; over the past years, such an environment was available as stand-alone modules running typically on large main frames; as a result, the grade of flexibility and the level of interaction were poor; nowadays, these limitations can be completely removed, thanks to the availability of fast and powerful DSP's combined with a developer-oriented interface.

SEGMENT DEFINITION

As known from the relevant literature, different choices can be made on the type of segments to be used; for example one could choice diphones, triphones, syllables, morphemes and at least whole words. In any case, regardless of the type of segment, a common problem exists: where exactly to cut the segments in order to be concatenated with the adjacent ones (preceding and following) and how to vary its time duration when the same segment appears at the beginning or at the end of a word.
In other terms this would mean that when extracting the generic segment, the developer should be able to mark in a clear way several characteristic points in the segment to be used subsequently as information for the concatenation.
For example, in our system, the basic segments are diphones, and we found that, besides the indication of the points where a segment starts and ends, it is also needed at least another pair of markers to indicate where to realize the concatenation (when it is required), and these markers depend on the phonetic context.
It is clear that all these markers have to be assigned to each segment, very often iteratively; as the total number of segments increases as a function of the phonetic complexity of the segment itself, a manual procedure is unthinkable.

SYSTEM DESCRIPTION

Hardware

The digital signal processing capabilities are demanded to a specific hardware designed at the same laboratory; it consists of a PC compatible add-on board. Actually the board can support different kind of processing, therefore it is not uniquely adopted for text to speech; in fact, it is

also capable to run speech coding techniques (ADPCM, RELP, LPC) and speech recognition algorithms based on a DTW approach.
Most of the DSP computations necessary for the algorithms mentioned are carried on a TEXAS Instruments TMS 320C25; the board includes other two processors, a MOTOROLA 68000 as CPU and an ALCATEL proprietary custom chip.
The communication between the DSP processor and the 68000 is performed via a shared memory (32 Kword), and all programs to be executed on the DSP are downloaded from the mass memory of the PC; therefore, after the initialization the DSP board acts as a stand alone system. Moreover, since the board is capable to directly interface a telephone line, the A/D and D/A functions and filtering are obtained with an industry standard CODEC; as a consequence the sampling frequency is fixed and equal to 8 kHz.

Software

The complete system can be considered as the combination of SW modules running on the PC under MS-DOS, DSP SW modules running on the board and executed by the TMS 320C25 and a supervisor module running on the board on the 68000 processor.
All SW modules running on the PC are written in Turbo PASCAL ver. 4.0 while the basic DSP modules have been written directly in TMS 320C25 Assembly in order to optimize the timing of crucial routines. Moreover, one of the main characteristic of this software modules is their intrinsic flexibility that offers the developer several options ranging from the choice of an alternative A/D and D/A board up to even more detailed parameters like the number and type of markers to be used for the concatenation. This is a very essential feature, since the amount of knowledge and information one can put in the segmentation process is inherently dependent on the expertise of the developer; with our system the level of details is not fixed, it can be increased and updated at any time.

SOFTWARE TOOLS

Acquisition, segmentation and coding

Generally, the segments necessary for a concatenative speech synthesis are extracted from suitable sentences uttered by a professional speaker; once the speaker's voice has been recorded as audio signal on convenient storage media (open reel tape, DAT), the developer has to convert the audio signal in data files: this stage of the development is called the acquisition; it basically consists in sampling the audio signal and store it as PCM coded files. The developer, using our speech workstation, has several choices, he may use the CODEC on the DSP board or other A/D boards available in the PC; it is also foreseen the possibility that the acquisition process will be performed elsewhere, in this case PCM files will be already available and the user has just to download them.
The segmentation will be realized using the PCM data files; the developer has the facility to display on the screen the waveform of the signal and, at the same time, to listen to the entire sentence or to a portion of it identified by a couple of cursors; to facilitate the task, additional information can be displayed on the screen as for example a formant tracking computed in real time.
On the basis of these information and iteratively listening to the acoustic feedback, the designer can select the portion of speech signal he wants to extract; the segment will be identified by a starting frame and an ending frame; further information, as for instance, the concatenating frames identified by another couple of markers, will notify which frame of the segment has to be used for concatenation, depending upon the phonetic context.
The user can utilize up to ten markers, whose meaning and usage can be defined in a set-up menu.
All these information will be stored in an appropriate header in order to be used by the subsequent stage that is the coding; this module performs the LPC coding of the relevant portion of the PCM files on the basis of the information retrieved from the corresponding header and will repeat this process for all the PCM files and all the headers.