A FUZZY APPROACH TO TEXT-TO-SPEECH SYNTHESIS

Enzo Mumolo, William Costanzo
Dipartimento di Elettrotecnica, Elettronica ed Informatica
Universita' di Trieste
via Valerio 10, 34127 Trieste, Italy
Ph/Fax: +39.40.676.3861/3460
E-mail: mumolo@univ.trieste.it

Abstract

In this paper we describe a fuzzy approach for the synthesis of a speech waveform from a phonetic description in Italian language. The system is based upon a set of fuzzy rules which linguistically describe the transitions between phonemes. The fuzzy system has several interesting properties, such as an easy management of the set of rules and the possibility to continuously improve the system by adding more knowledge to it using linguistic descriptions. Some experimental results are included, and future possible developments are outlined.

1 Introduction

The most popular approaches for performing TTS are the synthesis by concatenation and the synthesis by rule [2]. The fuzzy system described in this paper belongs to the latter category. In such algorithms, some issues are of concern: a detailed phonetic transcription is required at the input, which requires a linguistic analysis of the input text whose difficulty is greatly dependent on the language considered, a system of rules must be derived, which reflect our knowledge of the speech process and, finally, the quality of the resulting signal depends also on the computational model of speech.

Usually, phonemes are classified in terms of manner and place of articulation. The manner of articulation is concerned with the degree of constriction imposed by the vocal tract on the airflow, while place of articulation refers to the location of the most narrow constriction in the vocal tract. The following six categories of the manner of articulation have been considered in this work:

- vowel, in which air flows throw the vocal tract without constrictions
- liquid, similar to the vowels but that use the tongue as an obstruction
- nasal, which is characterized by a lowering of the velum, allowing airflow out of the nostril
- fricative, which employ a narrow constriction in the vocal tract which introduces turbulence in the air flow
- plosive, involving a complete closure and subsequent release of a vocal obstruction
- affricate, which is a plosive followed by a fricative.

Using the manner of articulation, the phonemes can be divided into broad categories. The detailed, finer discrimination of phonemes can be instead given by introducing the place of articulation, for which the following twelve categories have been used: anterior, open, rounded, voiced, bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental, velar. Furthermore, an 'any' feature has been added in this list for the definition of rules concerning transitions which do not depend on any phoneme. Using manner and place of articulation, any phoneme can be fully characterized in binary form [3]. However, a certain degree of imprecision is involved in this characterization, which thus should be fuzzy rather than strictly binary. For example, it may be that the /b/ phoneme, classically described as plosive, bilabial and voiced, involve also a certain degree of anteriority and rounding, as well as some other features.

The fuzzy rules, which describe the transitions between couple of adjacent phonemes, express linguistically the relation between the phonetic features and the speech model parameters. As such, the fuzzy system yields a simple and compact, yet accurate, description of the signal. Knowledge about speech production can be quite easily formalized and added to the fuzzy rules using linguistic expressions, thus opening for the development of simpler TTS systems with the possibility to reach higher levels of quality.

The flow chart is represented in Fig.1; it must be noted that this work is concerned only with the signal generation, thus no suprasegmental information has been given to the resulting signal.

This paper is organized as follows. In Section II the low-end processing of the system is outlined. In Section III the fuzzification, defuzzification and fuzzy rules are described. In Section IV some typical experimental results are reported. Final remarks and future developments are reported in Section V.

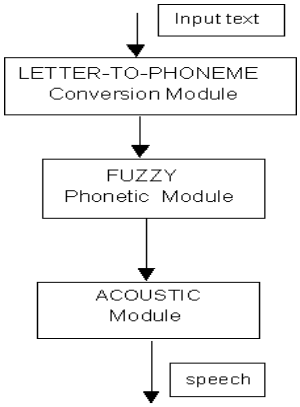

Figure 1: Structure of the conversion process.

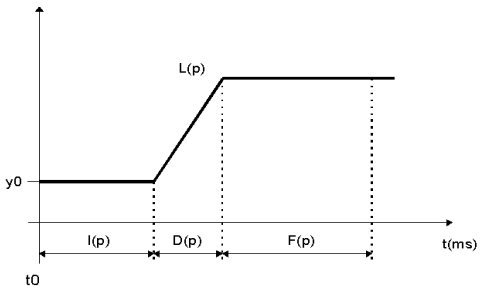

Figure 2: Profile of the synthesis parameter 'p'.

2 Letter to phoneme conversion

This conversion is performed using a table look-up approach, since in Italian language there is an almost 1:1 correspondence between graphemic and phonemic symbols. However, the following points have to be noted. Since the fuzzy rules are not concerned with the production of /C1C2/ sounds, a discontinuity symbol - $ - is introduced and included to produce consonants pairs. The discontiuity is a small pause with a given duration which can be controlled. An additional phoneme - /SIL/ - is introduced to represent the initial or final sections of the sentences.

3 The Acoustic Module

The synthesis of the output speech is performed using a reduced Klatt formant synthesizer [4]. This system is basically composed by a parallel filter bank for the vocal tract modeling in presence of voiceless source and a cascade of filters for the vocal tract modeling for voiced sounds. It is controlled by fifteen parameters, namely the first three formants and bandwidths, the bypass AB, the amplitude AV for voiced sounds and the amplitudes AF, AH and A2F-A6F for the fricative noise generator, updated every 5 ms. Since the fuzzy rules, however, describe the transitions between couples of phonemes from a dynamic point of view, a model of the parameters profiles has been introduced. The profile of each synthesis parameter 'p' is described with four control features, namely the initial and final intervals I(p) and F(p), the duration D(p) and the locus L(p), as reported in Fig.2. The I(p) control feature determines the starting point of the transition, whose slope and target values are given by the D(p) and L(p) features. The parameter holds the value specified by their locus for an interval equal to F(p) ms; however, if other parameters have not completed their dynamic, the final interval F(p) is prolonged.. The I(p), F(p), and D(p) parameters are expressed in milliseconds, while the target depends on what synthesis control parameter is involved; for example, for frequencies and bandwidths the locus is expressed in Hz, while for amplitudes in dB.

The output of the phonetic module, described in the next Section, is given in terms of such four parameters; of course the translation to the synthesis parameters required by the synthesizer must be performed.

4 The Fuzzy Phonetic Module

This module, which is the central part of the system, reads the phonetic string by couples of adjacent symbols and computes the control parameters I(p), F(p), L(p) and D(p) for each of the fifteen synthesis parameters. Nevertheless, not all the phonetic transitions are directly treated by the fuzzy rules: only the vowel to consonant (/VC/) transitions and consonant to vowel (/CV/) transitions are considered. This approximation is perceptually motivated; in fact only in those cases a real phonetic transition occurs, while the consonant to consonant transitions (/C1C2/) can be considered rather as a discontinuity. The degree of this approximation depends on the actual consonants considered but it is generally acceptable. Clearly, this approximation greatly simplifies the development of the rules. The generation of the /C1C2/ sounds is performed introducing an additional phoneme, called /$/, which is implemented as a pause with no sound generation and thus introduce a discontinuity.. The generation of a /C1C2/ sound, therefore, is split-up into two transitions, namely /C1$/ and /$C2/, managed by the fuzzy rules as usual. Another additional phoneme, called /SIL/, is introduced for representing the silence at the beginning and at the end of the phrases.

For simplicity, only a subset of the phonemes in Italian language were considered in this work. This subset is sufficient for achieving a complete intelligibility of a general text. Their classification in terms of the manner of articulation is as follows: vowels, liquids, nasals, fricatives, plosives, affricates.

Clearly, all the quantities involved, namely phonemes and control parameters, are fuzzified, as described in the following of the paper.

4.1 Phoneme and Control Parameters Fuzzification

As reported in the Introduction, the phonemes are classified into broad classes by means of the manner of articulation; then, the place of articulation is assigned to them. Therefore, each phoneme is described by an array of nineteen articulatory features, six of them are boolean variable and represent the manner of articulation and the remaining thirteen are fuzzy and represent the place of articulation. In this way, the phonetic description appears as an extension of the classical binary definition described for instance by Fant in [3], and a certain vagueness in the definition of the place of articulation of the phonemes is introduced. Representing the array of features as follows:

(vowel, plosive, fricative, affricate, liquid, nasal | any, rounded, open, anterior, voiced, bilabial, labiodental, alveolar, prepalatal, palatal, vibrant, dental, velar)

the /a/ phoneme, for example, can be represented by the array:

$$[1,0,0,0,0,0|1,0.32,0.9,0.12,1,0,0,0,0,0,0,0,0] \quad (1)$$

indicating that /a/ is a sonor vowel, with a degree of opening of 0.9, of rounding of 0.32, and it is anterior at a 0.12 degree. Similarly, the /i/ phoneme can be described by:

$$[1,0,0,0,0,0|1,0,0.06,0.9,1,0,0,0,0,0,0,0,0.1] \quad (2)$$

The /b/ phoneme, on the other hand, can be considered a plosive sonor phoneme, bilabial and slightly velar, and therefore it can be represented by

$$[0,1,0,0,0,0|1,0,0,0,1,1,0,0,0,0,0,0,0.2] \quad (3)$$

The two arrays reported as an example, have been partitioned for indicating the boolean and the fuzzy fields respectively. Such arrays, defined for each phoneme, are the membership values of the fuzzy articulatory features of the phonemes. On the other hand, the I,D,F and L fuzzy variables, defined in a continuous universe of discourse, can take any value in their interval of definition. The fuzzy sets for these variables have been defined as follows:

- Duration D(p). The global range of this fuzzy variable is 0-130 ms, with trapezoidal membership functions as shown in Fig.3. Fuzzy values:

  Very_Short, Medium_Short, Short, Medium, Medium_Long, Long, Very_Long

  In fig.3 such values are indicated as (B3, B2, B1, M, A1, A2, A3).

- Initial Interval I(p). As D(p), this fuzzy variable is divided into a 0-130 ms interval. The fuzzy values (cfr. fig.3) are, in this case:

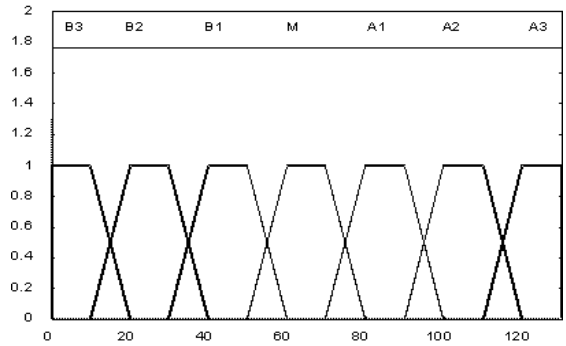

Figure 3: Membership functions of D(p), I(p), F(p). Horizontal axis in ms.

Instantaneous, Immediate, Quick, Medium, Medium_Delayed, Delayed, Very_Much_Delayed

- Final Interval F(p). The numeric range is 0-130 ms and the fuzzy values are the same as shown in Fig.3.

- Locus L(p). The fuzzy values of this variable depend on the actual variable to be controlled. For AV, AH and AF the fuzzy values are:

  Zero, Very_Low, Low, Medium_Low, Medium, Medium_High, High, Very_High

  and their membership functions are equally distributed between -12 and 80 dB with the same shape of Fig.3. The other gain factors, namely A2F-A6F and AB, can take one of the following values:

  Very_Low, Low, Medium_Low, Medium Medium_High, High, Very_High (4)

  and the range is 0-80 dB in the same manner as before.

  The values of L(F1), L(F2) and L(F3) are named as in (4), with trapezoidal membership functions uniformly distributed from 180 to 1300 Hz, 550 to 3000 Hz and 1200 to 4800 Hz for the first, second and third formants respectively. Finally, the loci of the bandwidths B1, B2 and B3 can take one of the names described in (4), and their membership functions are regularly distributed as in fig.3 on 30 - 1000 Hz for B1, 40 - 1000 Hz for B2 and 60 - 1000 Hz for B3.

4.2 Fuzzy Rules and Defuzzification

By using linguistic expressions which combine the above linguistic variables with fuzzy operators, it is possible to formalize the knowledge about the phoneme transitions for speech generation. In general, the rules involve the

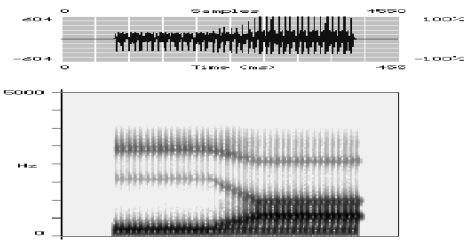

Figure 4: Spectrogram of the /ia/ transition.

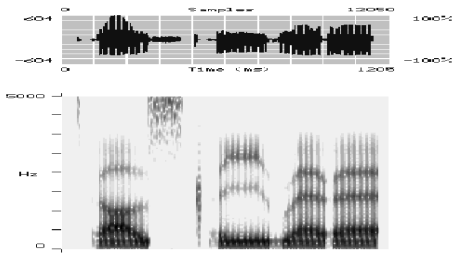

Figure 5: Spectrogram of the synthetized word /traskrivere/.

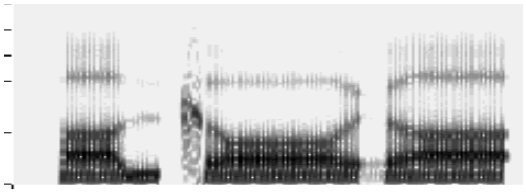

Figure 6: Spectrogram of the synthetized word /anko'ra/

actual and the future phonemes; thus, in these last rules, each of the expressions appears in the form: *If actual phoneme Is Any AND target phoneme ...* where the vocalic phonemes membership to the 'Any' variable is equal to 1.

Moreover, in general the fuzzy expressions involve the fuzzy operators AND, NOT and OR. Since the manner of articulation well partitions the phonemes in separated regions, the rules have been organized in banks, one for each manner. The rule decoding process is completed by the defuzzification operation, which is performed with the fuzzy centroid approach [5].

5 Discussion and Experimental Results

The fuzzy approach is summarized as follows:

- convert the input text in phonetic form
- for each pair of adjacent phonemes, activate all the fuzzy rules with the membership values associated to the phonemes
- perform deduzzification

As an example, let us suppose that the input string is /SIL//i//a/. The transitions are therefore the following: /SIL//i/+/i//a/+/a//SIL/. The rules pertaining to the vowel/vowel transitions are activated.

The spectrogram of the resulting sound is reported in Fig.4.We report, as another example, the conversion of the phonetic sequence /traskrivere/. The resulting spectrogram is shown in Fig.5, and the ability of the fuzzy system to generate plosive and vibrant transitions is evident. Another example is given in fig.6

6 Final Remarks and Conclusions

In this paper a novel approach for text-to-speech synthesis has been described. Many approximations - in the number of phonemes and in their membership to the articulatory features characterization, in the number of fuzzy rules and in the number of synthesis parameters - have been included for simlicity in this system. Nevertheless, the resulting speech has a good quality. From this point on, many improvements can be introduced in the system, basically by removing those approximations and by refining the fuzzy rules. Finally, the same approach could be used also for the development of speech analysis/synthesis algorithms for transmission and storage purposes. In this case, in fact, the transmitter should extract the place/manner of articulation feature from speech, code them and send them through a channel. The receiver reconstructs the original signal using the fuzzy algorithm.

In conclusion, the described approach is quite interesting because it yields TTS systems with much easier control of the generated quality and because it opens to many future research directions in several fields of speech and language engineering.

References

[1] Allen, J. , 'Synthesis of Speech from Unrestricted Text', *Proceedings of IEEE,* Vol.64, 1976, pp.422-433.

[2] Carlson R, Granstrom B. , 'A Text to Speech System Based Entirely on Rules', *ICASSP 76,* pp.686-688.

[3] Fant G. , 'Speech Sounds and Features', *The MIT Press, Cambridge, England*, 1973

[4] Klatt D.H. , 'Software for a cascade/Parallel Formant Synthesizer', *JASA, March 1980, pp.971-995.*

[5] Ralescu A.L. , 'Applied Research in Fuzzy Technology', *Kluwer, Academic Publisher, Boston, 1994.*