INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 5, 227–235, 2002
© 2002 Kluwer Academic Publishers. Manufactured in The Netherlands.

Improved Speech Synthesis Using Fuzzy Methods

DOINA JITCA
Institute of Information Science, Romanian Academy, Iasi Branch, 6600 Iasi, Romania

HORIA NICOLAI TEODORESCU
Romanian Academy and Technical University of Iasi, Romania
Hteodor@zeta.etc.tuiasi.ro

VASILE APOPEI AND FLORIN GRIGORAS
Institute of Information Science, Romanian Academy, Iasi Branch, 6600 Iasi, Romania

Abstract. The paper presents theoretical support for and describes the use of a fuzzy paradigm in implementing a TTS system for the Romanian language, employing a rule-based formant synthesizer. In the framework of classic TTS systems, we propose a new approach in order to improve formant trace computation, aiming at increasing synthetic speech perceptual quality. A fuzzy system is proposed for solving the problem of the phonemes that are prone to multi-definitions in rule-based speech synthesis. In the introductory section, we briefly present the background of the problem and our previous results in speech synthesis. In the second section, we deal with the problem of the context-dependent phonemes at the letter-to-sound module level of our TTS system. Then, we discuss the case of the phoneme /l/ and the solution adopted to define it for different contexts. A fuzzy system is associated with each parameter (denoted $F1$ and $F2$) to implement the results of the complete analysis of the phoneme /l/ behavior. The knowledge used in implementing the fuzzy module is acquired by natural speech analysis. In the third section, we exemplify the computation of the synthesis parameters $F1$ and $F2$ of the phoneme /l/ in the context of the two syllable sequences. The parameter values are contrasted with those obtained from the spectrogram analysis of the natural speech sequences. The last section presents the main conclusions and further research objectives.

Keywords: TTS system, letter-to-sound module, phoneme synthesis, formant parameters

1. Introduction

Recently, several priority research programs of the Romanian Academy addressed the use of computer science in relation to the Romanian Language. A notable outcome of our work in speech processing is the implementation of a Text-to-Speech (TTS) application for the Romanian language. We focused our research on achieving a good level of intelligibility and naturalness for the synthetic speech. In this respect, two aspects of the speech signal production were addressed, namely the glottal wave generator and the vocal tract modeling.

It is known that the glottal model makes a decisive contribution in speech signal generation (Fant et al., 1985; Fant and Lin, 1988), because the glottis is a nonlinear dynamic system that has a major influence on some important aspects of speech prosody. The glottal generator control we have used to obtain specific spectral features for the synthesized speech wave has been presented in Apopei et al. (2000).

Also an essential issue for the quality of the synthesized speech is the correct modeling of the vocal tract resonances. This actually relies on the accurate modeling of the resonances in speech generation. In this paper, we mainly discuss this second aspect, and

we present a novel method for obtaining an increased naturalness of the generated speech in the Romanian language.

Our approach to speech synthesis is based on the synthesis-by-rule concept (Klatt, 1987), in which speech is produced by matching phonetic and linguistic rules deduced from the analyzed speech data (phoneme sequences and prosodic features). The characterizations for target values (in the central part of the speech sounds) and transitions, as well as for pitch, amplitude and stress, are stored in rules; these rules are language-specific. Reliable and extensive knowledge of the speech production mechanism is essential in rule building (Grigoraş et al., 2000). We extracted this knowledge by analyzing a Romanian natural speech corpus.

Incompleteness, uncertainty, and complex dependencies characterize classical speech knowledge databases. For coping with such complex data, a good approach is to use artificial intelligence and, specifically, the fuzzy formalisms. The usefulness of the fuzzy approach to implement a speech production model was discussed in Grigoraş et al. (1999). The use of the natural language-like rules for the formant dynamics implementation was recommended in Raptis and Carayannis (1997), while the use of fuzzy-like rules for the pitch and spectrum has been suggested in Teodorescu et al. (1988).

Our research on nonlinearities and nonstationarity in speech production (Grigoraş et al., 1998, 1999) demonstrated the suitability of considering fuzzy logic for speech production modeling (Teodorescu et al., 1999).

Fuzzy dynamical sets (Kosanovic et al., 1996) were also employed in a study of Romanian phonemes recognition (Rodriguez et al., 2000), based on the assumption that fuzziness might cope successfully with the variability of human speech.

Modeling phenomena related to speech co-articulation effects and prosody requires rule finding and implementation. Finding rules involves parameter sequence computation (for modeling the glottis) and formant dynamics (for generating intelligible and natural sound sequences).

The paper presents theoretical support for and describes the use of fuzzy logic paradigms in implementing a rule-based synthesizer. The example presented deals with computing the lowest two formant traces for the liquid consonant /l/, considering the word context of the corresponding synthesized phoneme. The proposed model may be subject to generalization for co-articulation effects modeling.

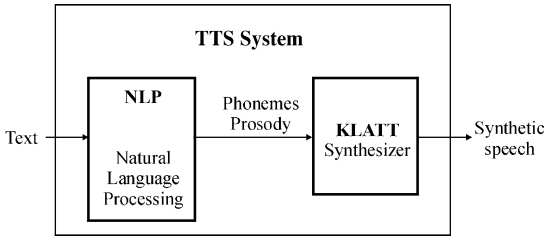

Figure 1. The diagram of a general TTS system.

2. The Letter-to-Sound Module for a Romanian-Language TTS System

Figure 1 represents a diagram of a TTS system based on a formant synthesizer. The output of the natural language processing (NLP) module consists of a sequence of phonemes and speech prosody elements for the currently synthesized speech segment (phoneme, word, or phrase). The sound generation module is a typical Klatt formant synthesizer. Its output is a time sequence of sound frames; during each frame the parameters for speech synthesis are considered to be constant.

The NLP module is composed of three submodules: the letter-to-sound transducer (LTS), the morpho-syntactic analyzer and the prosody generator (Dutoit, 1997). Regarding the LTS submodule, it is possible to organize the task using several strategies that generally fall into two classes, namely, dictionary-based and rule-based approaches. The second approach reduces the amount of information stored in the dictionary to a small number of exceptions and uses only a set of letter-to-sound rules for phonetic transcription.

Our TTS system translates the graphemes into the corresponding sounds only by means of rules that consider all the possible phonetic contexts of each grapheme (Jitcă et al., 2001). In the TTS application for the Romanian language, we have implemented rules of the following form:

"With G (the set of graphemes) and P (the set of phonemes), and given a grapheme $g \in G$, a left context $C_l \in G^n$ of the grapheme, and a right context $C_r \in G^n$ of the grapheme, $p \in P$ is a mapping of g and is written $g \Rightarrow P \mid C_l\_C_r$" (Wolters, 1997).

An entry in the phoneme definition table corresponds to each phoneme of the language. Under general circumstances, the definitions are sequences composed of synthesis parameter sets, suitable for the control of both stable and transitory phases of the phoneme. These parameter sets contain flags intended to assign

distinctive meanings to the parameters, when applied in synthesis.

The aforementioned definition of the phoneme, because it assumes fixed sets of parameters, is not satisfactory under all effective co-articulation contexts. Although the sound of the synthetic phoneme may be acceptable in some contexts, in others it may sound poorly, or, even worse, the output may sound closer to another phoneme. The occurrence of these situations lowers the synthetic speech intelligibility. On the other hand, providing several different definitions of the same phoneme for each context and increasing the number of rules induce a complex structure to the LTS module of the TTS application.

For the phonemes that imply different sets of parameters in various contexts, we introduced a hybrid definition, based on a fuzzy paradigm. Namely, we compute some of the parameter values for the synthesis, accordingly to each context. A specific flag is used to denote the phonemes with this particular treatment. When the flag is encountered during phoneme sequence translation, the system uses the fuzzy definition list to look up the phoneme, and the corresponding parameters are computed. The other parameters are given by the crisp values contained in the set relative to the phoneme's general definition.

In the next sub-sections, the case of the phoneme /l/ with two formant parameters (denoted $F1$ and $F2$) is presented. We will show that different parameter values are needed to synthesize the phonemes in different phonetic contexts. The method presented here for the first two formants ($F1$ and $F2$), i.e., the formants affecting the intelligibility of the phoneme), can be extended directly to the higher formants.

2.1. *Parameter Computation in Formant Synthesis by Means of a Fuzzy Paradigm*

In this section, we present a solution for assigning a fuzzy-set-based definition to the phoneme corresponding to the liquid consonant /l/. Usually, speech synthesis relies on expert knowledge extracted from acquired natural speech. The solution adopted for implementing our system is suggested by the well-known capability of the fuzzy paradigm to support the fusing of logic propositions, including qualitative attributes. The analysis was directed towards the study of the dynamics of acoustic parameter values. Analyzing the phoneme /l/ in various contexts, we derived the following rules for the target (i.e., to be currently synthesized) phoneme:

- For the currently synthesized phoneme, the central frequencies of the formants $F1$ and $F2$ are closely related to their values for the previous and the following phonemes;
- Values for the central frequency of the $F1$ formant range from 350 Hz to 450 Hz (male voice); and
- Values for the central frequency of the $F2$ formant range from 1200 Hz to 1600 Hz (male voice).

We propose a fuzzy system to implement the rules to control the two parameters (namely the central frequencies of the formants $F1$ and $F2$), based on the analysis of acquired natural speech data. Recall that a fuzzy system is defined by its variables, a set of rules, and the fuzzy operators (connectives) used. In this section, the term 'rule' refers hereafter to the rules of the fuzzy system.

Our TTS application uses two fuzzy systems for computing the central frequencies for $F1$ and $F2$. We use a Takagi-Sugeno fuzzy system model, with two input variables and one crisp output variable (Yager and Filev, 1994). To simplify notations, we denote the first two formants, as well as their central frequencies, by $F1$ and $F2$. The meanings result from the context.

The input variables are related to the values of the computed parameter given by the definitions of the previous and next phoneme. The crisp output of each fuzzy system provides the value of the parameters $F1$ and $F2$ for the currently synthesized phoneme /l/. We preferred a trapezoidal form for the membership functions, because the derived rules are operational on rather wide domains of the input variable. Both fuzzy systems are similarly defined (the same fuzzy sets with respect to number and shape); only the ranges of the membership functions are different (see Figs. 2 and 3).

The typical rule for the two fuzzy systems is described using the general statement:

$$\mathrm{L}^{(i)}\text{: IF } x_1 \text{ is } SF_1^i \quad \text{AND } x_2 \text{ is } SF_2^i \quad \text{THEN } y^i = c_1^i \cdot x_1 + c_2^i \cdot x_2,$$

where SF_j^i are the fuzzy sets of the input variables $x_j (j = 1, 2)$, c_j^i are real valued parameters, y^i is the crisp output for the rule $\mathrm{L}^{(i)}$, and $i = 1, 2, \ldots, M$.

The diagram of such a fuzzy system having two input variables and M rules is depicted in Fig. 2.

According to Fig. 2, the output of the system is the weighted mean of the outputs y^i and is given by the

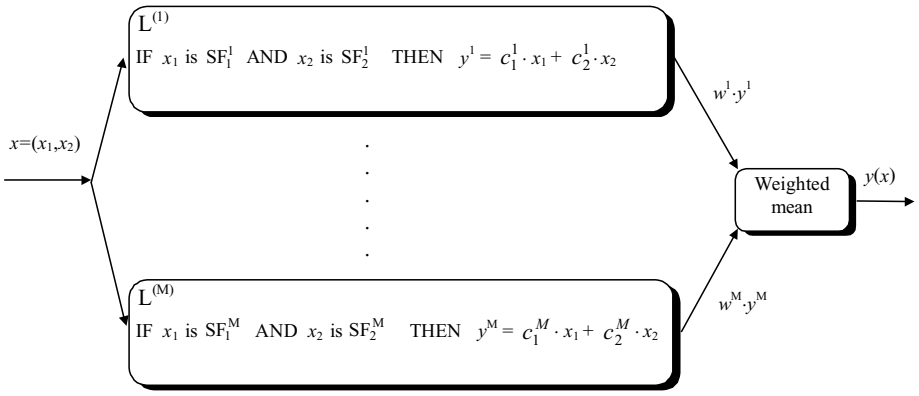

Figure 2. Takagi-Sugeno fuzzy system with two input variables and M rules.

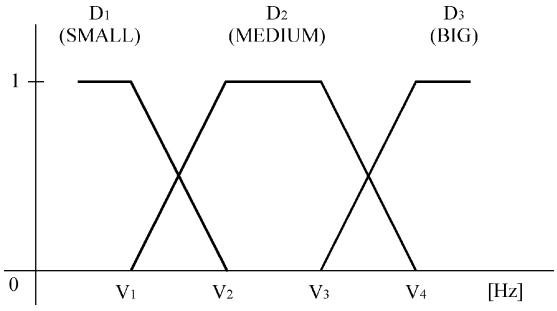

Figure 3. The generic membership functions for input variables x_1 and x_2.

Eq. (1):

$$y(x) = \frac{\sum_{i=1}^{M} w^i \cdot y^i}{\sum_{i=1}^{M} w^i} \tag{1}$$

The weighting makes use of the truth degree of the antecedent in the rule i. The weight w^i is computed as:

$$w^i = \prod_{j=1}^{2} \mu_{SF_j^i}(x_j), \tag{2}$$

where $\mu_{SF_j^i}(x_j)$ represents the membership degrees of the x_j values to the corresponding fuzzy sets SF_j^i. The application presented at the end of this section illustrates the computation of the truth degrees.

The universe of discourse is covered with three membership functions, D_1, D_2, D_3, named "SMALL", "MEDIUM", and "BIG", respectively (see Fig. 3).

The domains are delimited by the V_1, V_2, V_3, and V_4 frequency values that are specific to each input variable (x_1 and x_2).

2.2. *Definition of Variables for the Two Fuzzy Systems*

The central frequency of the $F1$ formant for the current phoneme, /l/, is the output of one system which has the inputs $F1$p (central frequency value of previous phoneme, x_1) and $F1$n (central frequency value of the next phoneme, x_2). The other system that computes the frequency value for the formant $F2$ is defined similarly (x_1 stands for $F2$p, and x_2 stands for $F2$n). The situation of the phoneme /l/ located in the first or in the last position of the word requires special treatment. Indeed, in the first situation there is no previous phoneme, while in the second case, there is no next phoneme. For these two situations we introduce the following rules:

- In the case of a leading /l/ in the currently synthesized word, the parameters $F1$p and $F2$p of the phoneme labeled as "previous" are taken to be the same as the parameters $F1$n and $F2$n of the "next" one, and
- In the case of an ending /l/, the parameters $F1$n and $F2$n of the phoneme labeled as "next" are taken to be the same as the parameters $F1$p and $F2$p of the "previous" one.

2.3. *The Membership Functions Corresponding to the Two Fuzzy Systems*

The membership functions for the input variables $F1$p and $F1$n are obtained by forcing a set of values for the generic annotations $V1$–$V4$ in Fig. 3, namely $V1 = 350$ Hz, $V2 = 400$ Hz, $V3 = 450$ Hz, and $V4 = 500$ Hz. Using these values, the input membership functions for the first fuzzy system are drawn in Fig. 4.

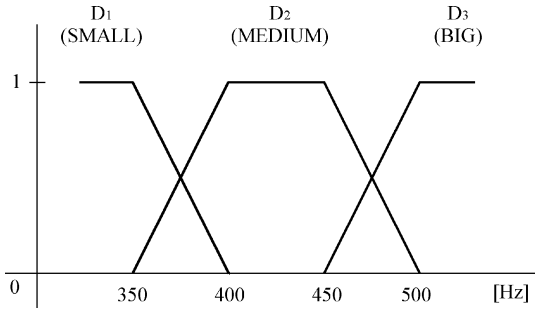

Figure 4. The membership functions for the variables $F1\text{p}$ and $F1\text{n}$.

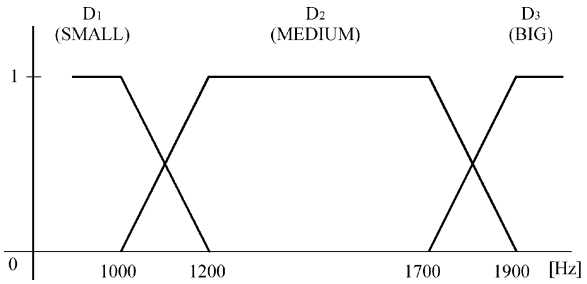

Figure 5. The membership functions for the variables $F2\text{p}$ and $F2\text{n}$.

In the same manner, the membership functions for the input variables $F2\text{p}$ and $F2\text{n}$ are obtained by forcing the values $V1 = 1000$ Hz, $V2 = 1200$ Hz, $V3 = 1700$ Hz, and $V4 = 1900$ Hz. For these values, the input membership functions for the second fuzzy system are drawn in Fig. 5.

Thus, the values of the parameters $F1$ and $F2$ can be assigned to one of the categories "SMALL", "MEDIUM" and "BIG" that represent the fuzzy sets D_1, D_2 and D_3, respectively.

2.4. The Rules

Table 1 presents the rules of the fuzzy system computing the central frequency of the $F1$ formant. The rules for the fuzzy system that computes the central frequency for the $F2$ formant are presented in Table 2.

Table 1. The rules of the fuzzy system that computes the central frequency of the $F1$ formant.

1.	IF $F1_\text{p}$ is D_1 AND $F1_\text{n}$ is D_1 THEN $y = F1 = 0.5 \cdot F1_\text{p} + 0.5 \cdot F1_\text{n}$
2.	IF $F1_\text{p}$ is D_1 AND $F1_\text{n}$ is D_2 THEN $y = F1 = 0.5 \cdot F1_\text{p} + 0.5 \cdot F1_\text{n}$
3.	IF $F1_\text{p}$ is D_1 AND $F1_\text{n}$ is D_3 THEN $y = F1 = 1.0 \cdot F1_\text{p} + 0.0 \cdot F1_\text{n}$
4.	IF $F1_\text{p}$ is D_2 AND $F1_\text{n}$ is D_1 THEN $y = F1 = 0.5 \cdot F1_\text{p} + 0.5 \cdot F1_\text{n}$
5.	IF $F1_\text{p}$ is D_2 AND $F1_\text{n}$ is D_2 THEN $y = F1 = 0.75 \cdot F1_\text{p} + 0.25 \cdot F1_\text{n}$
6.	IF $F1_\text{p}$ is D_2 AND $F1_\text{n}$ is D_3 THEN $y = F1 = 0.75 \cdot F1_\text{p} + 0.25 \cdot F1_\text{n}$
7.	IF $F1_\text{p}$ is D_3 AND $F1_\text{n}$ is D_1 THEN $y = F1 = 0.00 \cdot F1_\text{p} + 1.00 \cdot F1_\text{n}$
8.	IF $F1_\text{p}$ is D_3 AND $F1_\text{n}$ is D_2 THEN $y = F1 = 0.75 \cdot F1_\text{p} + 0.25 \cdot F1_\text{n}$
9.	IF $F1_\text{p}$ is D_3 AND $F1_\text{n}$ is D_3 THEN $y = F1 = 450$ Hz

Table 2. The rules of the fuzzy system that computes the central frequency of the $F2$ formant.

1.	IF $F2_\text{p}$ is D_1 AND $F2_\text{n}$ is D_1 THEN $y = F2 = 1200$
2.	IF $F2_\text{p}$ is D_1 AND $F2_\text{n}$ is D_2 THEN $y = F2 = 0.0 \cdot F2_\text{p} + 1.0 \cdot F2_\text{n}$
3.	IF $F2_\text{p}$ is D_1 AND $F2_\text{n}$ is D_3 THEN $y = F2 = 0.5 \cdot F2_\text{p} + 0.5 \cdot F2_\text{n}$
4.	IF $F2_\text{p}$ is D_2 AND $F2_\text{n}$ is D_1 THEN $y = F2 = 1.0 \cdot F2_\text{p} + 0.0 \cdot F2_\text{n}$
5.	IF $F2_\text{p}$ is D_2 AND $F2_\text{n}$ is D_2 THEN $y = F2 = 0.5 \cdot F2_\text{p} + 0.5 \cdot F2_\text{n}$
6.	IF $F2_\text{p}$ is D_2 AND $F2_\text{n}$ is D_3 THEN $y = F2 = 0.75 \cdot F2_\text{p} + 0.25 \cdot F2_\text{n}$
7.	IF $F2_\text{p}$ is D_3 AND $F2_\text{n}$ is D_1 THEN $y = F2 = 0.5 \cdot F2_\text{p} + 0.5 \cdot F2_\text{n}$
8.	IF $F2_\text{p}$ is D_3 AND $F2_\text{n}$ is D_2 THEN $y = F2 = 0.75 \cdot F2_\text{p} + 0.25 \cdot F2_\text{n}$
9.	IF $F2_\text{p}$ is D_3 AND $F2_\text{n}$ is D_3 THEN $y = F2 = 1600$ Hz

3. A Case Study

The performance of our system is illustrated by computing the parameters for two syllable sequences, /la-le-li/ and /li-lo-lu/, followed by the synthesis of the above-mentioned sequences. The central frequencies of the formants $F1$ and $F2$ are derived for the consonant /l/, taking into consideration the context, namely the "previous" and the "next" phonemes. For each sound sequence, we analyzed the acquired natural speech utterances, in terms of the central frequencies of the $F1$ and $F2$ formants, and then we performed the synthesis of both sequences. Figure 6 depicts the spectrogram for the sound sequence /la-le-li/.

The traces of the formants $F1$ and $F2$ for the three /l/ phonemes are marked and are annotated on the

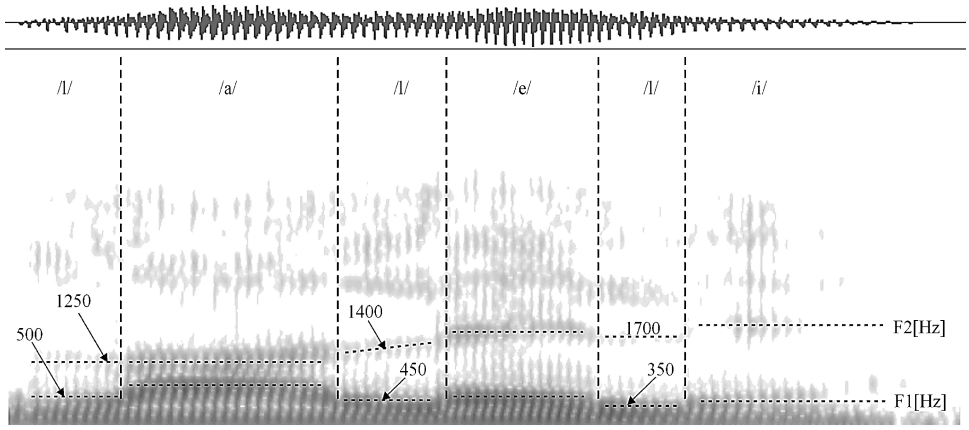

Figure 6. Spectrogram of the acquired syllable sequence /la-le-li/. Upper side: time-domain signal.

spectrogram (Fig. 6). The corresponding formant values are 500, 450, and 350 Hz for $F1$, and 1250, 1400, and 1700 Hz for $F2$. These values must be compared with those obtained for the synthesized /l/ phonemes. Precisely, the absolute values are not essential, but the ratios of the current formant value to the value of the formant of the previous and next phonemes are of interest. The method for computing the $F1$ and $F2$ traces for a more natural-like synthesis is explained in detail in this section.

According to the convention previously introduced to manage the exceptions of the leading /l/ phoneme context, $F1$p equals $F1$n and $F2$p equals $F2$n for this consonant at the beginning of /la-le-li/ sequence. For the three /l/ phonemes, the values for $F1$ and $F2$ of the previous and next phonemes are specified as follows:

- For the first /l/ phoneme, $F1_p = 700$ Hz, $F1_n = 700$ Hz, $F2_p = 1250$ Hz, and $F2_n = 1250$ Hz;
- For the second /l/ phoneme, $F1_p = 700$ Hz, $F1_n = 500$ Hz, $F2_p = 1250$ Hz, and $F2_n = 1800$ Hz; and
- For the third /l/ phoneme, $F1_p = 500$ Hz, $F1_n = 350$ Hz, $F2_p = 1800$ Hz, and $F2_n = 2000$ Hz.

For the first phoneme, the antecedents are:

- $F1_p$ is D_3 AND $F1_n$ is D_3 with a truth degree equal to 1, causing only Rule 9 to fire, with the result that $y^9 = 450$ and $F1 = 1 \cdot y^9 = 450$ Hz, and
- $F2_p$is D_2 AND $F2_n$ is D_2, with a truth degree equal to 1, causing only Rule 5 to fire, with the result that $y^5 = 0.5 \cdot 1250 + 0.5 \cdot 1250 = 1250$ and $F2 = 1 \cdot y^5 = 1250$ Hz.

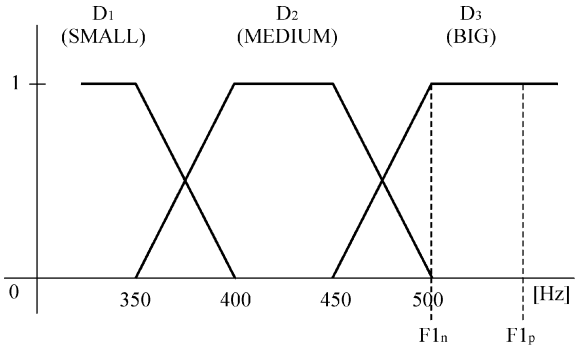

Figure 7. The input variables $F1_p$, $F1_n$ for the second /l/ phoneme, marked on the universe of discourse.

For the second /l/ phoneme, the antecedents are:

- $F1_p$ is D_3 AND $F1_n$ is D_3 with a truth degree equal to 1, causing only Rule 9 to fire, with the result that $y^9 = 450$ and $F1 = 1 \cdot y^9 = 450$ Hz (see Fig. 7);
- $F2_p$ is D_2 AND $F2_n$ is D_2 with a truth degree equal to 0.5, causing the firing of Rule 5, that is, $y^5 = 0.5 \cdot 1250 + 0.5 \cdot 1800 = 1525$; and
- $F2_p$ is D_2 AND $F2_n$ is $D_3$with a truth degree equal to 0.5, causing the firing of Rule 6, that is, $y^6 = 0.75 \cdot 1250 + 0.25 \cdot 1800 = 1388$.

The fuzzy system for the $F2$ parameter combines the outputs of Rules 5 and 6 to generate the output value (see Fig. 8) $F2 = 0.5 \cdot y^5 + 0.5 \cdot y^6 = 0.5 \cdot 1525 + 0.5 \cdot 1388 = 1456$ Hz.

For the third /l/ phoneme, the antecedents are:

- $F1_p$ is D_3 AND $F1_n$ is D_1 with a truth degree equal to 1, causing only Rule 7 to fire, with the result that

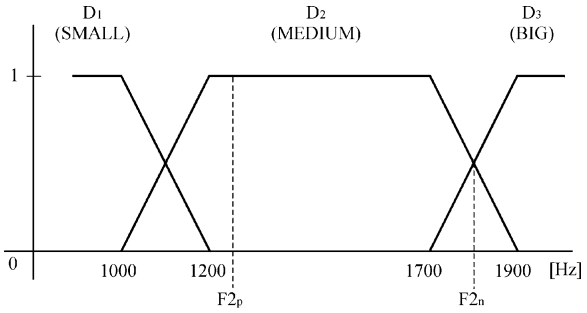

Figure 8. The input variables $F2_p$, $F2_n$ for the second /l/ phoneme, marked on the universe of discourse.

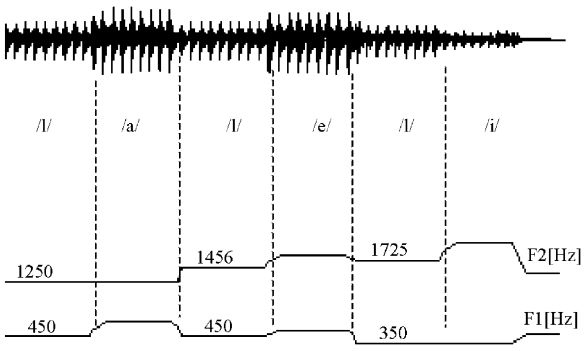

Figure 9. Computed formant traces for /la-le-li/ sequence and synthetic speech output. Upper side: time-domain signal.

$y^7 = 0 \cdot 500 + 1 \cdot 350 = 350$ and $F1 = 1 \cdot y^7 = 350$ Hz;

- $F2_p$ is D_2 AND $F2_n$ is D_3 with a truth degree equal to 0.5, causing the firing of Rule 6, that is, $y^6 = 0.75 \cdot 1800 + 0.25 \cdot 2000 = 1850$; and
- $F2_p$ is D_3 AND $F2_n$ is D_3 with a truth degree equal to 0.5, causing the firing of Rule 9, that is, $y^9 = 1600$.

The fuzzy system for the $F2$ parameter combines the outputs of rules 6 and 9 to yield the output

$$F2 = 0.5 \cdot y^6 + 0.5 \cdot y^9$$
$$= 0.5 \cdot 1850 + 0.5 \cdot 1600 = 1725 \text{ Hz}.$$

Formant traces computed by the described fuzzy systems are depicted in Fig. 9, together with the time domain waveform of the synthetic speech.

The second part of this case study illustrates the operation of the fuzzy paradigm in computing the central frequencies of the $F1$ and $F2$ formants for the synthesized syllables sequence /li-lo-lu/. The spectrogram of the natural utterance /li-lo-lu/ is shown in Fig. 10.

The formant values are 350, 425, and 370 Hz for $F1$, and 1700, 1400, and 1200 Hz for $F2$. The value 1400 Hz for the second phoneme /l/ is a mean value of parameter $F2$ in transition from 2000 Hz to 900 Hz. We adopted this value to represent the value for the stable period because the transition cannot be reproduced in synthesis.

The same requirements for the exceptions regarding the context of the /l/ consonant at the beginning of the /li-lo-lu/ sequence induce $F1$p to be equal to $F1$n and $F2$p to be equal to $F2$n. The other values for the central frequencies of the $F1$ and $F2$ formants are specified as follows:

- For the first /l/ phoneme, $F1_p = 350$ Hz, $F1_n = 350$ Hz, $F2_p = 2000$ Hz, and $F2_n = 2000$ Hz;
- For the second /l/ phoneme, $F1_p = 350$ Hz , $F1_n = 500$ Hz, $F2_p = 2000$ Hz, and $F2_n = 850$ Hz; and
- For the third /l/ phoneme, $F1_p = 500$ Hz , $F1_n = 370$ Hz, $F2_p = 850$ Hz, $F2_n = 1000$ Hz.

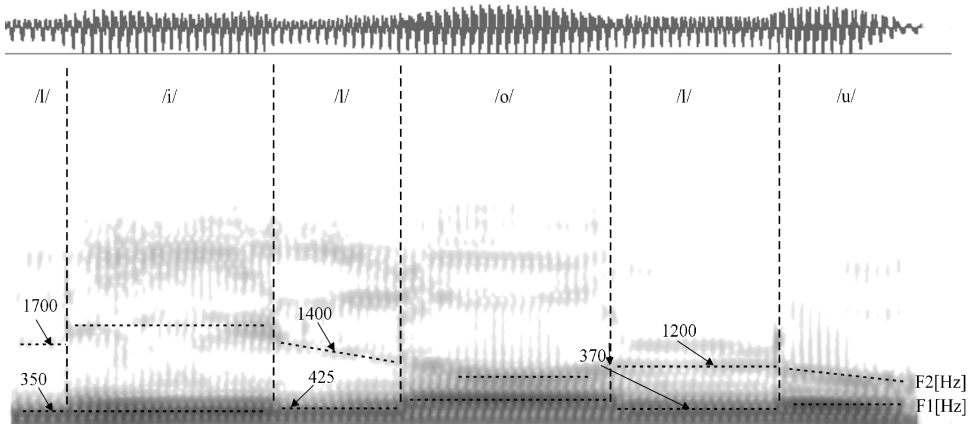

Figure 10. Spectrogram of the acquired syllable sequence /li-lo-lu/. Upper side: time-domain signal.

Subsequently, we present the calculations for the antecedents and the rules fired for the second sound sequence, /la-lo-li/.

For the first /l/ phoneme, the antecedents are:

- $F1_p$ is D_1 AND $F1_n$ is D_1 with a truth degree equal to 1, causing only Rule 1 to fire, with the result that $y^1 = 0.5 \cdot 350 + 0.5 \cdot 350 = 350$ and $F1 = 1 \cdot y^1 = 350$ Hz, and
- $F2_p$ is D_3 AND $F1_n$ is $D_3$with a truth degree equal to 1, causing only Rule 9 to fire, with the result that $y^9 = 1600$ and $F2 = 1 \cdot y^9 = 1600$ Hz.

For the second /l/ phoneme, the antecedents are:

- $F1_p$ is D_1 AND $F1_n$ is D_3 with a truth degree equal to 1, causing only Rule 3 to fire, with the result that $y^3 = 0.5 \cdot 350 + 0.5 \cdot 500 = 425$ and $F1 = 1 \cdot y^3 = 425$ Hz, and
- $F2_p$ is D_3 AND $F1_n$ is D_1 with a truth degree equal to 1, causing only Rule 7 to fire, with the result that $y^7 = 0.5 \cdot 2000 + 0.5 \cdot 850 = 1425$ and $F2 = 1 \cdot y^7 = 1425$ Hz.

For the third /l/ phoneme, the antecedents are:

- $F1_p$ is D_3 AND $F1_n$ is $D_1$with a truth degree equal to 1, causing only Rule 7 to fire, with the result that $y^7 = 0 \cdot 500 + 1 \cdot 370 = 370$ and $F1 = 1 \cdot y^7 = 370$ Hz, and
- $F2_p$ is D_1 AND $F1_n$ is $D_1$with a truth degree equal to 1, causing only Rule 1 to fire, with the result that $y^1 = 1200$ and $F2 = 1 \cdot y^1 = 1200$ Hz.

The formant traces $F1$ AND $F2$ of the synthetic sequence /li-lo-lu/ and the time domain waveform are shown in Fig. 11.

The two examples of implementation with fuzzy systems, as presented above, demonstrate the capabilities of the method. Indeed, the $F1$ and $F2$ values can be adapted conveniently to all the context situations and result in a more intelligible synthetic speech. The examples also indicate that fuzzy systems represent an efficient solution for implementing small knowledge bases associated with different features of natural speech, in order to incorporate them in the synthesis procedure. The main benefit consists of generating synthesis parameter values of the phonemes close to those in natural speech. We performed a large number of experiments using this strategy (based on the fuzzy paradigm), and

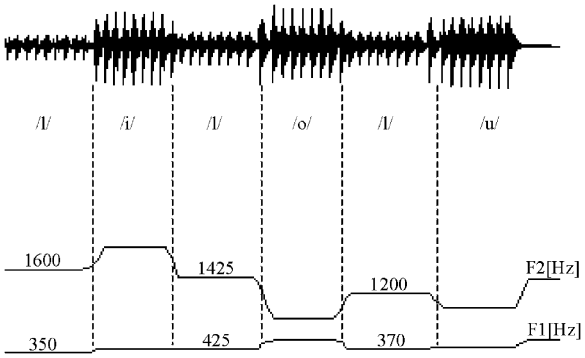

Figure 11. Computed formant traces for /li-lo-lu/ sequence and synthetic speech output. Upper side: time-domain signal.

all showed that the perceptual quality of the synthetic speech is increased considerably.

The use of the fuzzy control allows us to model of the natural variability of the central frequency of the formants and is simpler and more intuitive then other methods. Instead of fuzzy systems, we could use polynomial interpolators or neural networks. However, polynomial equivalent models (5 to 12 polynomial coefficients) are more intensive computationally then fuzzy systems. Using an artificial neural network may be equally complex and does not allow the direct interpretation of the results by a speech engineer.

4. Conclusions and Further Work

Our experiments in TTS synthesis demonstrated that classic formant synthesizers are very efficient and well suited for the Romanian language if supplied with enough language-specific information. The synthesis quality is satisfactory in most cases. However, in some cases when the features of the sounds were highly context-dependent, we had to modify the standard rules in two ways: we altered the crisp definition of the sounds (Iles and Simmons, 1994), and we added an intelligent module to improve the computation of the formant traces. The results presented are based on the manual extraction and are dependent on expert's knowledge. At this moment we have no specially developed tool for the automatic extraction of the variability curve of the central frequency of the formants, and the operation must be executed by the human expert.

Also, regarding the generic capabilities of formant synthesizers, our experiments showed that further improvements are required, including a variable length for

phones and greater flexibility in using the fundamental frequency for stress assignment. Moreover, better implementation of aspiration and frication are needed. Nevertheless, the theoretically imperfect structure of formant synthesizers (source-filter model) compels the user to make further empirical adjustments of some parameters (amplitude of voicing, patterns and degrees of frication and aspiration).

We conclude that the proposed approach based on the fuzzy paradigm leads to an increased intelligibility for a formant synthesizer and is a viable alternative in order to advance to the next generation of high quality rule-based speech synthesizers. In our Romanian-TTS system, empirical crisp calculus, previously used in speech synthesis applications, is replaced with a flexible fuzzy control. The fuzzy control aims to obtain correct target values for the parameters of the classical Klatt synthesizer (Klatt, 1980), as required for the desired high perceptual quality of the generated speech. The study of the speech production mechanism and the formalism of artificial intelligence and fuzzy techniques were applied to obtain the rule base and to infer accurate instant values for the parameters of the speech synthesizer. The current implementation may be linked to further developments focusing on better modeling of the articulatory apparatus. Moreover, the same approach may be used to make the synthesis environment-adaptable, according to rules similar to those presented in Teodorescu et al. (1988).

A further research objective is to incorporate the data-mining approach to derive fuzzy rules. Indeed, there is a huge amount of data to be analyzed and, probably, a large number of rules to be applied to achieve high quality speech synthesis. The manual analysis of large databases is an almost impossible task, and there is no guarantee that the rules derived are optimal. The most appealing data-mining approaches for this task seem to be the evolutionary techniques; among them, genetic programming seems to be best suited to the task.

Acknowledgments

The authors acknowledge the support of a research grant from the Romanian Academy during 2000 and 2001. Thanks are due to the three anonymous referees and to the Special Issue Editor for many helpful comments.

References

Apopei, V., Jitcă, D., Grigoraş, F., and Teodorescu, H.N. (2000). Naturalness in speech synthesis by fuzzy control of the glottal parameters. *IIZUKA'2000 Conference CD-ROM Proceedings*. Fukuoka, Japan: Fuzzy Logic Systems Institute.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four parameter model of glottal flow (Research Report STL-QPSR 4, KTH). Stockholm, Sweden: Royal Institute of Technology, pp. 1–13.

Fant, G. and Lin, Q. (1988). Frequency domain interpretation and derivation of glottal glow parameter (Research Report STL-QPSR 2-3, KTH). Stockholm, Sweden: Royal Institute of Technology, pp. 1–21.

Grigoraş, F., Apopei, V., Jitcă, D., and Teodorescu, H.N. (2000). Conclusions from a research on soft-computing rule-based speech synthesis for Romanian language. *ECIT2000 CD-ROM Proceedings*. Iaşi, Romania: Coda Press.

Grigoraş, F., Teodorescu, H.N., and Apopei, V. (1998). Nonlinear analysis and synthesis of speech. *Studies in Informatics and Control*, *7*(1):57–72.

Grigoraş, F., Teodorescu, H.N., Jain, L.C., and Apopei, V. (1999). Fuzzy and knowledge-based control for speech synthesis. *ECC'99 CD-ROM Proceedings*. Karlsruhe, Germany: VDI/VDE Geselschaft.

Iles, J. and Ing-Simmons, N. (1994). Rsynth_2.0, Text-to-Speech software, ftp://svr-ftp.eng.cam.ac.uk comp.speech/sources.

Jitcă, D., Apopei, V., and Grigoraş, F. (2001). Text to speech system for Romanian language based on formantic synthesis. Communicated paper, Iaşi Academy Days, Romania (unpublished paper).

Klatt, D. (1980). Software for cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*:971–995.

Klatt, D. (1987). Review of Text-to-Speech conversion for English. *Journal of Acoustic Society of America*, *82*:737–793.

Kosanovic, B.R., Chaparro, L.F., and Sclabassi, R.J. (1996). Signal analysis in fuzzy information space. *Fuzzy Sets and Systems*, *77*:49–62.

Raptis, S. and Carayannis, G. (1997). Fuzzy logic for rule-based formant speech synthesis. *EUROSPEECH'97 Proceedings*. Rhodes, Greece, vol. 3, pp. 1599–1602.

Rodriguez, W., Teodorescu, H.N., Grigoraş, F., Kandel, A., and Bunke, H. (2000). A fuzzy information space approach to speech signal non-linear analysis. *International Journal of Intelligent Systems*, *15*(4):343–363.

Teodorescu, H.N., Apopei, V., Grigoraş, F., Jitcă, D., Nica, D., and Buzatu, O. (1999). Formant speech synthesis improvement by AI and fuzzy methods (Research Report). Bucharest: Romanian Academy (unpublished paper).

Teodorescu, H.N., Chelaru, M., Sofron, E., and Adascalitei, A. (1988). Adaptive speech synthesis. *ITG-Fachbericht 105, Digitale Sprach-verarbeitung—Prinzipien und Anwendungen*. Berlin:VDE-Verlag GmBh, pp. 183–188.

Wolters, M. (1997). A diphone-based Text-to-Speech system for Scottish Gaelic. A Master Thesis presented to the University of Bonn, Bonn, Germany.

Yager, R.R. and Filev, D.P. (1994).*Essentials of Fuzzy Modeling and Control*. New York: Wiley.