FUZZY LOGIC FOR RULE-BASED FORMANT SPEECH SYNTHESIS

S. Raptis and G. Carayannis
Speech Synthesis Team,
Institute for Language and Speech Processing,
22 Margari St., 115 25, Athens, Greece.
Tel. +30 1 6712250, Fax: +30 1 6741262, E-mail: spy@ilsp.gr

ABSTRACT

Fuzzy set theory and fuzzy logic has been initiated by Zadeh back in 1965 to permit the treatment of vague, imprecise, and ill-defined knowledge in an concise manner. One of the unique advantages of fuzzy logic is that it is capable of directly incorporating and utilizing qualitative and heuristic knowledge in the form of causal if-then production rules for reasoning and inference. On the other hand, rule-based speech synthesis based on formants makes considerable use of rules for numerous of the tasks it involves, e.g. graphemic to phonemic transcription, coarticulation, concatenation, and duration rules etc. These rules also take the if-then form with their antecedent (condition) part describing the context of the rule and their decedent an appropriate action to be taken. The main motivation for introducing fuzzy logic in the synthesis-by-rule paradigm, is its ability to host and treat uncertainty and imprecision both in the condition part of the rule as well as its decedent part. This may be argued to significantly reduce the number of required rules while rendering them more meaningful and human-like.

1. INTRODUCTION

Probably the main source of information used for speech synthesis, results from analysis of natural speech. We seek for rules describing the changes that are observed to the target values of an isolated segment, when this segment appears within different contexts. A plethora of such heuristic rules can be found in literature, especially concerning acoustic phonetics and spectrographic analysis. Such rules are usually easier to express in qualitative rather than numerical form. It is, for example, known that [1]:

“There is a clear downward movement at then end of the second and third formants of a vowel when a bilabial nasal [m] follows.” (see Fig. 1)

“These is a comparatively small movement of formants at then end of a vowel when it is followed by an alveolar nasal [n].” (see Fig. 1)

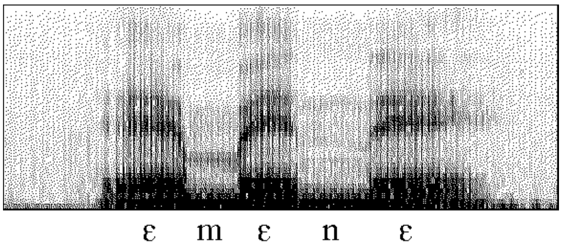

Figure 1. Spectrogram of the word [ɛmɛnɛ]

Another important source of information has its roots in the physiology of the human speech production system, i.e. its articulation mechanisms and characteristics. It is, for example, well known that an articulator that is not involved in the primarily articulation of a sound will take up or tend towards the articulation of the following sound. This phenomenon, known as anticipatory articulation, is the main reason for the nasalization of vowels (see Fig. 1).

It is argued that all this valuable information cannot be directly utilized using “conventional” rule format or when it can it requires numerous rules to represent it. On the other hand, fuzzy rules have inherent capabilities for treating vague and imprecise information of this sort providing the designer with expressiveness and flexibility.

2. “CONVENTIONAL” RULEBASES AND THEIR WEAKNESSES

Although such observations are quite valuable for synthesis purposes, they cannot be directly incorporated into a “conventional” segmental rulebase used in typical rule-based formant speech synthesizers, at least not until the context of each rule has been formally described and expressions like “comparatively small” are replaced by precise numerical values.

Even if we assume that these rules are converted in a strict numerical form, some problems may emerge when multiple of such rules are activated simultaneously for a specific segment, the main cause of them being their "relative strength". E.g. if one of the active rules suggests that the value of the second formant, F2, should be increased by 200Hz while another suggests that it should come to be 1500 Hz, what should be value attained by F2? It is important to note that the fact of more than one rule being active at a specific moment, certainly does not imply poor rule selection. On the contrary, this fact is expected to increase the generalization capabilities and the quality of the interpolation exhibited by the rulebase.

To effectively handle simultaneous rule activation, each rule should be provided with a strength. A rule's strength (which may be variable) captures the rule's relevance at the specific circumstances, the rule's applicability, or whatever the rulebase designer considers appropriate.

Among the most widely used rule formats are the SPE-like formats (introduced by Chomsky and Halle [2]) involving a focus, one or more actions, and a context specification:

focus → actions / context specification

for example, the hypothetical rule:

m → F2 - := 200 / {ɑ/ɔ/ɒ/o}[nonseg] ---

denotes that the value of the second formant of /m/ should decrease by 200 when appearing in the position of '---' in the context specification of the rule ('[nonseg]' denotes the set of all non-segmental symbols). Such rules can be considered to be of the form [3]:

if C then A

where C denotes the condition part and contains the focus and context specification, while A denotes the action to be taken. Using this format the above rule may be rewritten as:

if focus is /m/
and previous symbol is any of {ɑ/ɔ/ɒ/o}
and there is a word break
then F2 -:= 200

The condition part of this rule is a logical operation between binary valued predicates, each of which reduced to a boolean and thus the whole condition part reduces to a boolean. For this kind of rules, if it is desirable to provide a rule strength, this will have to be imposed externally. Such a strength is quite necessary when more than one of such rules may "fire" simultaneously, so as to determine the rule with the dominant influence. Clearly, the overall output of the rulebase for the focus symbol, will be a weighted average of the partial outputs of each of the active rules, for each of the parameters under control.

Common extension to the above format allow segment families (rather than unique segments) to be involved in the condition predicates which are defined based on descriptions of their articulation, e.g.:

if focus is /m/
and previous is (vowel) and (back) and (not high)
and there is a word break
then F2 -:= 200

This format improves to a degree the rule's expressiveness but renders the demand for handling multiple simultaneous activation stronger.

From the discussion above, it is clear that what is really missing is:

- A framework that will be able to take advantage of heuristic rules as extracted from spectrographic analysis and physiology of the human speech production system and utilize them in a rule-based synthesis scheme. This expressiveness should be available for both the rule condition (context specification) and the rule action.
- An efficient way of merging the partial actions of several rules that are simultaneously activated.

3. FUZZY SET THEORY

3.1. Fuzzy Sets

The heart of fuzzy theory is centered around the definition of a *fuzzy set*. In classic (crisp) set theory, a set is defined on a universe of discourse (domain), say U, through a characteristic function which assigns 1 or 0 to the elements of U thus discriminating between members and non-members of the set. Allowing partial membership to the set, i.e. permitting the characteristic function to obtain all the intermediate values between 0 and 1, the definition of a fuzzy set and its corresponding *membership function* emerges. Fuzzy sets provide the means for capturing non-binary concepts, which are most commonly found in practice.

For example, consider the case of a vowel. Vowels may be classified based on the position of attained by the tongue during their articulation: from front to back and from low to high. Phoneticians usually define these vowel features by drawing them on a chart having the first formant on the vertical axis and the difference of the second to the first formant on the horizontal, the origins residing at the upper right corner. In this chart, higher vowels tend to be placed upper and more front vowels tend to be placed more to the left. In this two dimensional space, say U, a crisp-style definition of the set of front vowels, would result in a set of the form depicted in Fig. 2a:

$$\text{front} = \{(x, y) \in U \mid x > 800\text{Hz}\}$$

while a fuzzy-style definition would probably be of the form depicted in Fig. 2b.

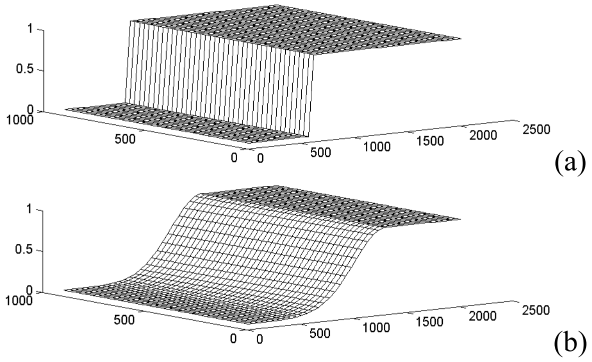

Figure 2. *Sets capturing "front" vowels, (a) crisp set, (b) fuzzy set.*

Crisp sets are considered to be a special cases of fuzzy sets. It is apparent that not only can we create a fuzzy set to capture front vowels but we can additionally define degrees of "frontness". So we may say that the degree of membership of /i/ in the set of front vowels is 1. while for /ɛ/ it drops to 0.8 ending up to 0.3 for /u/ and to 0 for /o/. Thus, when a more continuous scale than the one permitted by the binary framework of distinctive features is preferred, fuzzy sets are an excellent candidate.

For representing fuzzy sets, various shapes can be used for their membership function, e.g. Gaussian bells, trapezoids, triangles, etc. Figure 3 shows how one can define fuzzy sets for high, mid-high, mid, mid-low, and low vowels based on the value of their first formant.

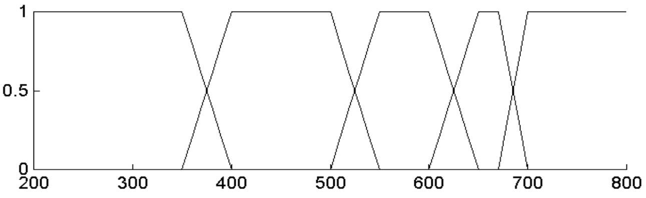

Figure 3. *Fuzzy sets for high, mid-high, mid, mid-low, and low vowels based on the value of their first formant*

Most crisp operations on sets may be directly ported to the fuzzy framework. Thus, we may acquire the union, intersection, and complement of fuzzy sets by generalizing the respective crisp set operations.

Although defining a set to capture the tongue position proved to be easy, this is not true for other articulatory characterizations. For example, it is not straightforward or even meaningful to define a fuzzy set to capture the family of stop consonants. It is however intuitively appealing to use fuzzy sets for the places of articulation. Defining a fuzzy set for the placement or even the rounding of the lips does may sense and, in fact, may be used to define very useful rules as explained below. Also appealing appears to be the definition of fuzzy sets upon the domains of appropriately chosen distinctive features of the phonemes.

3.2. Fuzzy Rules

Having defined fuzzy sets one may proceed to the definition of *fuzzy rules*. A fuzzy rule is, in effect, an implication between the rule's condition and action. A typical (Mamdani-type) fuzzy rule has the general form:

IF *x*1 IS *A*1 AND *x*2 IS *A*2 AND ...
THEN *y*1 IS *B*1 AND *y*2 IS *B*2 AND ...

where *xi* denotes the *i*-th input and *yj* denotes the *j*-th output which are both supposed to be fuzzy sets, and *Ai* and *Bj* are fuzzy sets defined on the same universe of discourse as *xi* and *yj* respectively. An example of a rule might be:

IF *focus_position* is FRONT and *next* is BILABIAL
THEN dF2 is NEGATIVESMALL

This rule accepts two inputs, namely the focused and the next symbol, and produces one output, which is the change that needs to be performed on the value of the second formant of the focused phoneme. FRONTVOWEL, BILABIAL, and NEGATIVESMALL are supposed to be fuzzy sets defined on appropriate universes of discourse.

Fuzzy rules deal with fuzzy sets, so when a numeric input needs to be supplied to the rule, a *fuzzification* process needs to take place. Similarly, the results output by the fuzzy rulebase are fuzzy sets and thus to retain a numeric output value a *defuzzification* process in required.

The condition part of the rule consists of predicates that do not reduce to boolean but to fuzzy sets, indicating the consistency of the specific input with the preconditions of the rule. Since fuzzy theory is in many of its aspects a generalization of crisp set theory, a boolean-like predicate may be considered as a special case of a fuzzy predicate.

The more relevant is a rule, the higher the resulting fuzzy sets will be. Performing an AND operation on these sets, is equivalent to calculating their disjunction. The result directly reveals the relevance of the specific rule to the current circumstances. If the focused symbol is indeed a front vowel then the rule is relevant and will fire strongly, while for center vowels it will still fire but in much lesser degree. The fire level of the rule proportionally affects the rule's output level.

Thus, fuzzy rules have an automatic mechanism for determining their relevance and acting accordingly. Moreover, the approximate reasoning theory efficiently merges the outputs of all the rules in a rulebase to produce the overall system output. This takes place by calculating the conjunction of all the partial results which is also a fuzzy set.

CONCLUSIONS

We believe that such rules may directly incorporate physical constraints of the vocal tract concerning the position and motion of the articulators across successive phonemes and that using such a form we may manage to both reduce and render more meaningful the rules required for high quality speech synthesis. Such rules might be easier to develop and maintain and might even be more straightforward to port to other languages.

Another considerable advantage of the use of fuzzy logic for rule-based speech synthesis stems from the quite extensive work that has already been (and is still being) carried out on *adaptive fuzzy systems*. Introducing a learning component in the fuzzy rule-base, one may adjust and fine tune the system behavior based on numerical data extracted by, for example, analysis of natural speech. This unique advantage of adaptive fuzzy systems, namely their ability to host both qualitative and numerical knowledge under the same framework, might prove to be quite exceptional.

REFERENCES

[1] P. Ladefoged, "A Course in Phonetics", Harcourt Brace College Publishers, 1993.

[2] N. Chomsky and M. Halle, "The Sound Patterns of English", Harper and Row, New York, 1968.

[3] L. F. M. ten Bosch, "From Data to Rules", In *Talking Machines: Theories, Models, and Designs*, G. Bailly, C, Benoît, and T. R. Sawallis (Eds.), Elsevier Science Publishers, B. V., 1992, pp. 13-26.

[4] S. G. Tzafestas, "Fuzzy Systems and Fuzzy Expert Control: An Overview", *The Knowledge Engineering Review*, Vol. 9:3, 1994, pp. 229-268