Softcomputing approach to segmentation of speech in phonetic units

M.Malcangi

Abstract— Speech-To-Text and Text-To-Speech applications are essentially based on an effective separation of phonetic units, so the segmentation of uttered speech into phonetic units is a key processing task for successfully implementing speech recognition systems. Softcomputing methods demonstrate to be more effective than other methods due to the capability neural networks and fuzzy logic to be trained by expert. This work phonetic segmentation of uttered speech that separates vowels from consonants is based on a fuzzy logic inference engine tuned by an expert using speech features distribution. Only time-domain feature-extraction algorithms are applied to speech to extract features, so minimum computational cost was achieved. Fuzzy decision logic is used to infer about phonetic units separation point. A set of tests has been executed to demonstrate that this approach can be effective in separating phonetic units, while requiring minimal computing power and reducing system complexity.

Keywords— Fuzzy decision logic, pitch estimation, speech energy, **s**peech segmentation, speech analysis, speech recognition, speech synthesis, zero-crossing rate.

I. Introduction

SPEECH is a quasi-stationary signal, so it can be divided into segments with stationary features. Speech segments can be classified into macro typologies, such as voiced or unvoiced, or in micro typologies such as phones (phonemes and allophones).

Voiced speech consists of vowels and semi-vowels, and unvoiced speech consists of consonant speech segments. Voiced speech segments are generated by glottis when it is in the vibration state (pulsing at the pitch frequency). Unvoiced speech segments are generated by air freely and randomly flowing through glottis (opened).

Phonemes and allophones are voiced and unvoiced speech segments characterized by a different constriction of the vocal tract.

Speech analysis, speech recognition, and speech synthesis rely on certain key information embedded in uttered speech [3], [4], [5], [6], [7], [8], [9]. Such information includes pitch, formants, energy, and so forth. These features are not stationary for an entire utterance (i.e. a word). Therefore, a frame-based processing technique is currently used.

Speech analysis is based mainly on signal processing algorithms for frequency mapping such as Fourier Transform (FT) or algorithms for frequency tracking such as linear prediction (LP). Stationary of frequency information in speech segments processed by such algorithms is a primary requirement for successful application of frequency features measurements in speech recognition, speech identification, and speech synthesis.

In speech recognition application isolating stationary speech segments is useful when a phone-to-text (PTT) approach is to be implemented. Best matching of a phone (phoneme or allophone) can be achieved only if its feature extraction is concerning mainly the segment part with minimal coarticulation information.

In speech synthesis based on concatenation of speech segments (phonemes, diphones syllables, etc.) the automatic segmentation of utterance is necessary to build-up the speech data base required by the speech synthesizer. Voiced/unvoiced endpoints of uttered speech are to be identified to effectively drive the speech synthesizer to switch the glottal source to be processed by the parametrically controlled vocal tract simulation.

Frame-based speech-signal processing consists of separating a short portion of the whole uttered signal. The duration of the segment must be short enough to assume the speech features in such a frame to be stationary [4], [15] .

Features can be considered stationary for a phonemic segment of the speech but not for the coarticulation part of it. The framing process is asynchronous with respect to time sequencing phonetic units in uttered speech, so most of the information about speech features that is extracted by a fixed-length frame process is imprecise. As a result, speech-processing performance is significantly reduced.

To overcome this problem, the framing process needs to be synchronized with the time position of the phonetic units in the uttered speech, so only stationary data will be subjected to speech-feature extraction. To achieve synchronization, a smart segmenting process needs to be implemented to allow the time position of phonetic units in the uttered speech to be identified.

Over the last decade, considerable effort has been devoted to automatically segmenting speech into phonetic units [2], [10], [11], [14], [16].

Solutions have been sought both in terms of parametric (i.e. algorithmic) signal-processing methods (frequency-domain feature extraction with pattern matching) [13] and in terms of nonparametric (i.e. linguistic) signal-processing methods (mainly artificial neural networks) [1].

This work proposes a mixed-method approach to segmenting uttered speech into phonetic units. The approach simplifies the extraction of speech features obtained through parametric signal processing by applying only a few time-domain signal-processing algorithms and by using a simple fuzzy-logic inference engine to pattern-match the phonetic units.

Time-domain signal processing and non–parametric pattern matching of phonetic units endpoint detection have been implemented to implement speech recognition (speech-to-text), speech identification (voiceprint), and speech synthesis (text-to-speech) on low-end microcontrollers in deeply embedded applications.

II. Process Framework

To separate phonemic information in uttered speech, energy, zero-crossing rate, and pitch have been computed by a set of time-domain measurement algorithms and then combined by a fuzzy logic-based inference engine. This computing process is described below (fig. 1).

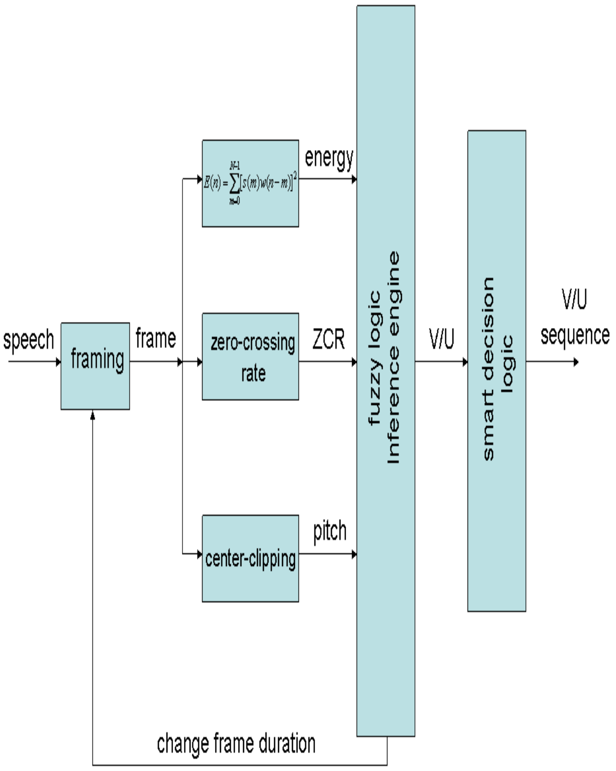

Fig. 2 process to segment speech in voiced/unvoiced segments

It is a frame-by-frame process that computes short-time energy, zero-crossing rate, and pitch rate. Short-time energy is combined with zero-crossing rate to classify voiced and unvoiced segments (V/U). On speech segments is executed also the measurement of pitch frequency. All the measurements are computed in time domain.

Energy, zero-crossing rate, and pitch are combined to infer about the phonetic class to which belong each segment (V/U). An higher smart decision logic layer streams the V/U decision to evaluate the end-points of the U/V segments. It recovers errors occurred at fuzzy logic decision layer considering the duration, considering pre and post decision.

A feedback has been introduced to automatically resize the frame duration to best match the endpoints of each phonetic unit. Frame resizing occurs when the fuzzy logic decision is not able to classify the current frame. This happens above all when a segment contains coarticulated voiced-unvoiced speech. Frame duration is then halved and the fuzzy logic decision is then applied separately to the frames.

A. Time-domain features computation

Speech has several crisp features related to its semantic content. Some of these are time parameters, such as amplitude, energy, and dynamics. Others are frequency parameters, such as pitch or formants.

Time-domain computation of speech features is computationally less intensive than frequency-domain computation. Time parameters can easily be computed with an across-the-board formula like

$$Q(n)=\sum_{m=0}^{N-1}T[s(m)]w(n-m) \qquad (1)$$

- *Q(n)* is the short-time calculation of a feature from a sampled audio signal *s(n)*
- *T* is a time-domain transformation function applied to signal *s(n)*, weighted by the window *w(n)*.

The *T* transformation is computationally easier than the one needed for frequency-domain transformation. The above formula can also be applied to compute frequency-domain signal features, such as pitch and zero-crossing rate.

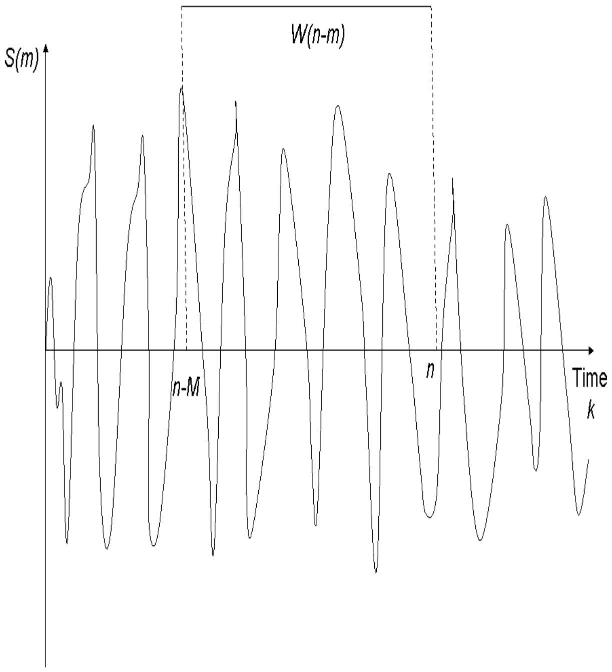

Fig. 2 a window of N samples in length superimposed on a speech sequence with only gradual variation

Windowing (fig. 2) serves to execute the short-time feature calculation, with such features treated as gradually variable information.

B. Short-time energy

Amplitude of speech signal is continuously variable with time. Root-Mean-Square (RMS) calculation can be executed on a windowed segment to represent such variability. Energy measurement is as representative as RMS, but it is computationally easer because square root calculation is avoided.

Short-time energy is computed applying the windowing technique. Energy is initially calculated using a 20 ms wide Hamming window, according to the following formula:

$$E(n) = \sum_{m=0}^{N-1} [s(m)w(n-m)]^2$$
$$w(m) = 0.54 - 0.46\cos(2\pi m/(N-1)), \text{for } 0 \le m \le N-1 \quad (2)$$
$$w(m) = 0, \text{otherwise}$$

The energy measure very effectively segments the speech signal into phonetic units, such as vowels and consonants, because the amount of energy is vastly different between two phonemes (fig. 3).

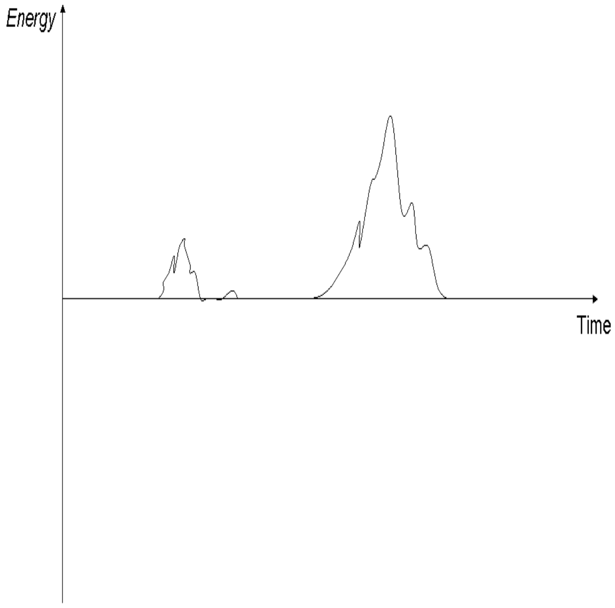

Fig. 3 energy is starkly different between two phonemes

The easily identifiable variation in energy enables to quickly distinguish voiced from unvoiced uttered speech.

Energy computed in large windows will fluctuate too slowly in time to obtain good information about phone nature. Energy computed in small windows is too rapidly varying and this can give false information about phone transition. Window duration can be adapted to the time dynamic of the uttered speech. Larger or shorter windows can be applied to segment speech so that transition effects due to coarticulation can be minimized.

C. Short-time zero-crossing rate

Zero-crossing rate (ZCR) is a synthetic measure of frequency characteristic of signal. It measures how many time the signal is crossing the zero-amplitude line in a unit time (window duration). Distribution of such frequency measurement is very indicative of the nature of a signal frame regarding its belonging to the class of quasi-periodic and the class of non-periodic signals.

The ZCR feature is calculated with the following formula:

$$ZCR(n) = \sum_{m=0}^{N-1} 0.5\left|sign(s(m)) - sign(s(m-1)\right|w(n-m)$$
$$(3)$$
$$sign(s(n)) = 1, \text{for } x(n) \ge 0$$
$$sign(s(n)) = \text{otherwise.}$$

A unit-gain rectangular window scaled by 1/N is applied to yield ZCR per sample. N is the length of the window.

Voiced and unvoiced phonemes have very distinctive ZCR measurements. These correlated closely to points of major energy concentration. The ZCR and energy measurements are clearly adequate to classify a sound as voiced or unvoiced [13].

D. Short-time pitch

Pitch is a frequency information related to the voiced/unvoiced nature of speech. Voiced speech (e.g. the vowels) embeds the glottis frequency. Unvoiced speech doesn't embed any pitch frequency. Such information, combined with short-time energy and short-time ZCR make more robust the voiced/unvoiced separation process.

Pitch is convolved with the vocal tract frequency response, so it need to be deconvolved. Such process is computationally intensive, and its computation cost is not comparable with the computation cost of energy and ZCR).

Pitch is estimated through center clipping technique. Such technique is applied to short-time frames of uttered speech. Compared to other techniques for estimating pitch, such as autocorrelation or cepstrum, this technique estimates pitch more precisely at lower computational cost [14].

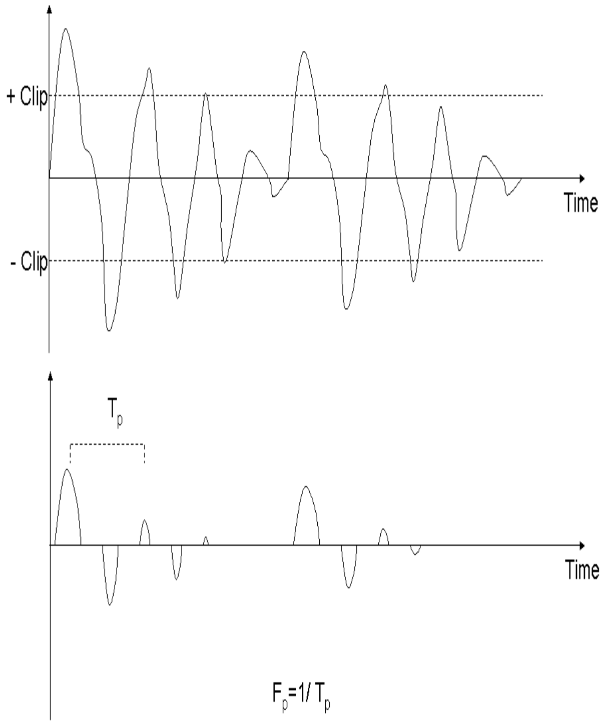

Fig. 4 detecting pitch by center-clipping a speech frame

Speech-frame center clipping (fig. 3) consists of setting a sample's low-amplitude levels to zero, while reducing its high-amplitude levels. This signal distortion is based on a clipping threshold tuned to about 30% of maximum signal amplitude.

Pitch measurement is useful to distinguish between voiced and unvoiced utterance frames. During transition from voiced to unvoiced speech frames, energy and ZCR measurements might not be sufficiently indicative of speech-frame unit type. In such cases, pitch measurement may be informative enough to correctly classify a frame as voiced or unvoiced.

III. Smart Decision Logic

A fuzzy-logic engine was tuned to infer speech classification in two distinct phoneme classes: vowels and consonants. The engine processes crisp measurements of energy, ZCR, and pitch. It then decides how to classify the frame. The inferred knowledge is based on a set of membership functions that depend on the distribution of input measurements and on a set of linguistic rules derived from experimentally observing the measurements that correlated to these two phonetic classes in the uttered speech. Voiced speech concentrates most of energy below 3 kHz. Unvoiced speech concentrate most the energy at higher frequencies.

Membership functions were modeled on the distribution of each measurement. Energy was modeled vis-à-vis the highest and lowest values, thus constructing a triangular fuzzification function. Voiced speech concentrates most of energy below 3 kHz. Unvoiced speech concentrate most the energy at higher frequencies (fig. 5).

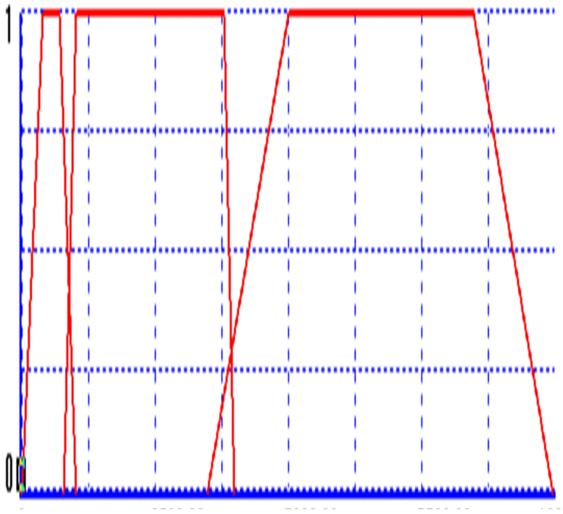

Fig. 5 membership functions to fuzzify short-time energy

Normalized ZCR distribution was directly transformed into a membership function by interpolating the distribution curve as a triangle or trapezoid (fig. 6).

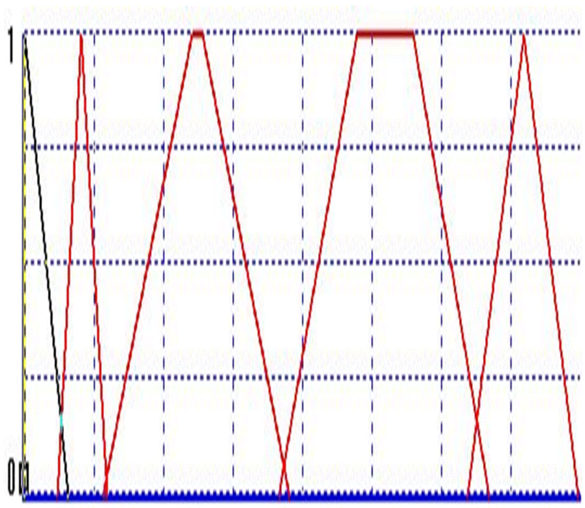

Fig. 6 membership functions to fuzzify short-time zero-crossing rate

Pitch intensity was mapped onto frequency so that it fuzzily discriminates between voiced and unvoiced speech units, as well as among different voiced phonemes.

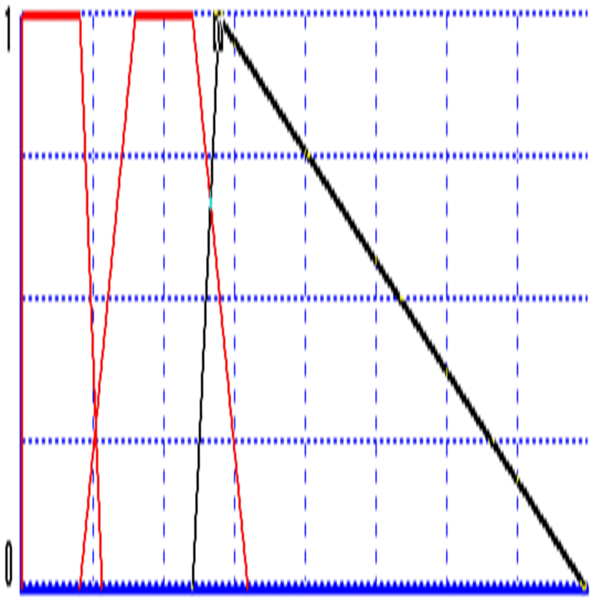

Fig. 7 membership functions to fuzzify short-time pitch

For each phoneme analyzed, rules regarding energy, zero-crossing rate, pitch, and membership in either a vocalic or a consonantal utterance unit were compiled.

IF STZCR IS Low
AND STE IS Average
AND STPITCH IS Average
THEN Segment IS Voiced

IF STZCR IS Low
AND STE IS High
AND STPITCH IS High
THEN Segment IS Voiced

IF STZCR IS VeryLow
AND STE IS High
AND STPITCH IS High
THEN Segment IS Voiced

Fig. 8 fuzzy rules to infer the voiced attribute of an uttered speech segment

A center of gravity method is then applied to fuzzify the final decision, so that, frame by frame, the uttered speech is segmented into vocoidal and contoidal speech units.

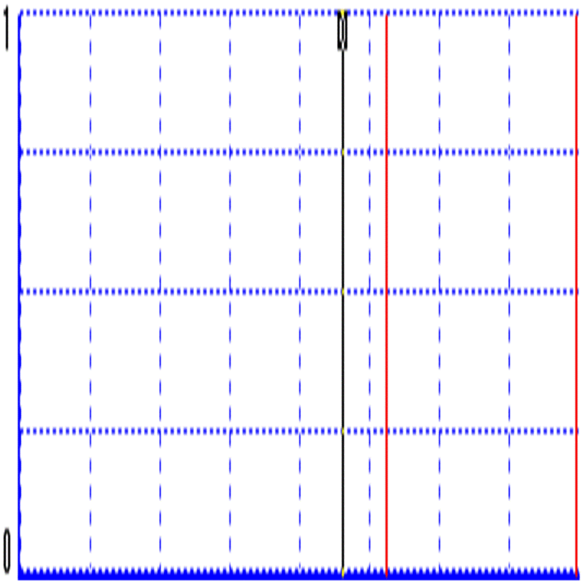

Fig. 9 fuzzy rules to infer the voiced/unvoiced attribute of an uttered speech segment

IV. Experimental results

To evaluate the performance of this approach to phonetically segmenting uttered speech, an experimental environment was modeled using MATLAB. The whole experimental system consists of a feature-extraction subsystem and a fuzzy-logic inference engine.

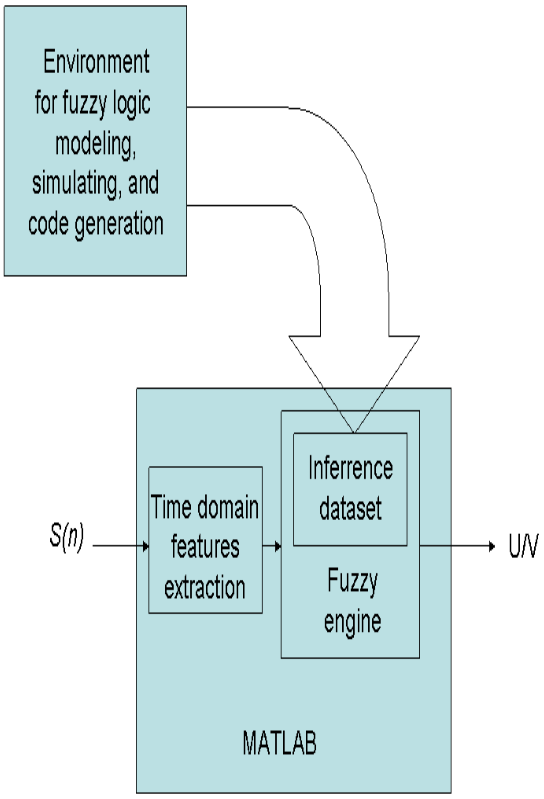
Environment for fuzzy logic modeling, simulating, and code generation
S(n)
Time domain features extraction
Inferrence dataset
Fuzzy engine
MATLAB
U/V

Fig. 10 experimental environment for smart segmentation of uttered speech running in MATLAB modeling environment

The inference engine was set up in a separate modeling environment (FUDGE) that enables the user to model membership functions and rules. Then, after simulation to fine-tune the inference engine, the configuration dataset is transferred to the MATLAB-coded engine that manages decision logic.

A set of tests was run on isolated words (the set of uttered digits from ZERO to NINE). Results for the segmented word SIX are shown in fig. 11

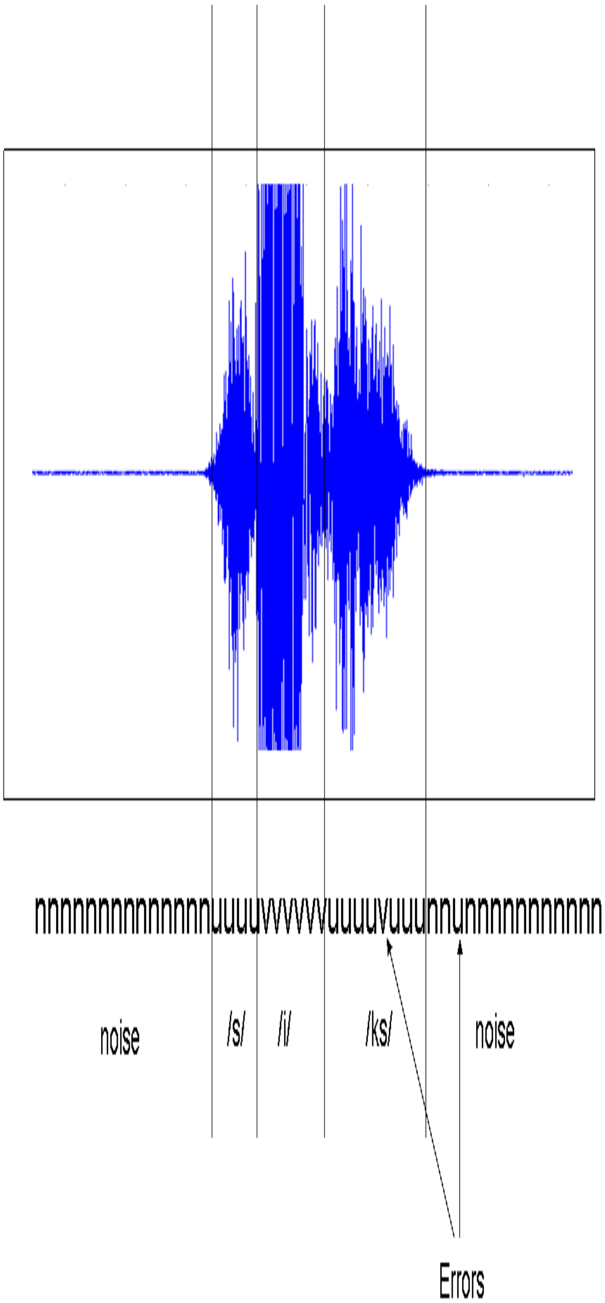

Fig. 11 sequence of segments of the uttered word SIX, showing starting and stopping points for the vowel *'i'* and the consonants '*s*' and '*x*': two wrong segments are also endpointed

The segmentation process demonstrate to be very effective. Some errors occur, but they are easily recoverable introducing a time duration criteria such as.

IF segment_duration
IS shorter THAN n_frames
AND
antecedent_frame = conseguent_frame
THEN segment BELONGS TO PREVIOUS

Such rule acts like a smoothing fuzzy filter on the symbolic sequence outputted by the segmentation engine. After this postprocessing, a correct segmentation has been achieved (fig. 12).

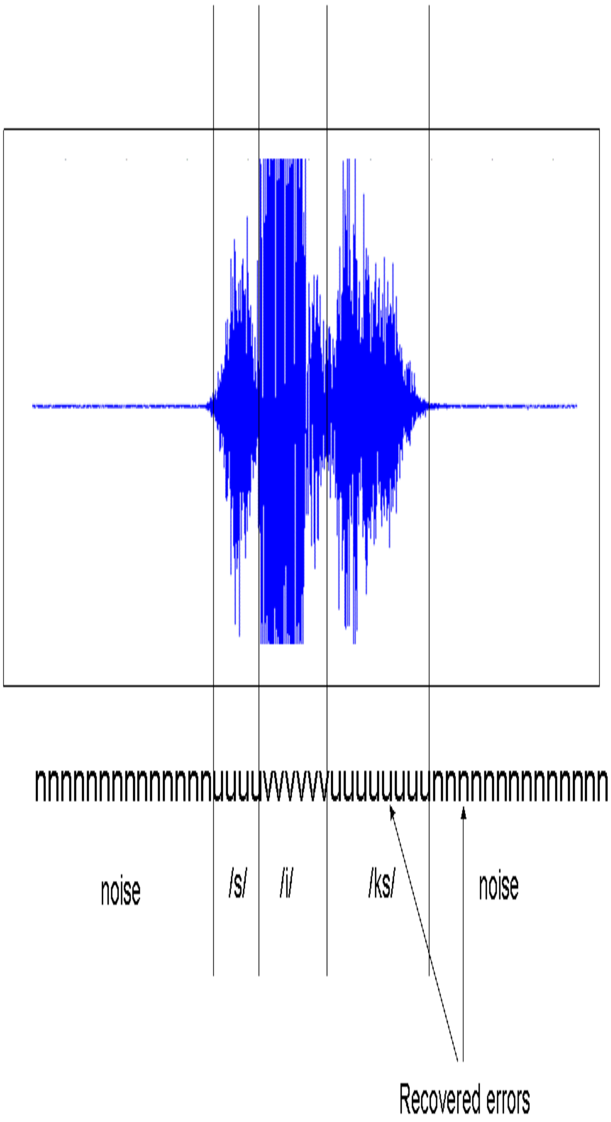

Fig. 12 sequence of segments of the uttered word SIX, showing starting and stopping points for the vowel *'i'* and the consonants '*s*' and '*x*': two wrong segments are also endpointed

V. Conclusion

Segmentation of speech in phonetic units is a very important task for successful implementation on most important speech application, including speech recognition, speech synthesis and, speaker identification. Because such applications are more and more required in embedded systems, smart processing approaches need to be investigated.

Time domain computation demonstrate to be effective at very low computational cost. Low computational cost can be achieved also using fuzzy decision logic because the fuzzy logic inferential engine can be run efficiently on a low-end microcontrollers.

The experimental results demonstrate that the combination of time-domain speech-feature measurement and fuzzy decision logic is simultaneously effective and very efficient. Computational complexity was kept to a minimum to allow low-cost implementation in deeply embedded systems for speech-based applications such as voice control.

Next steps of this research will be focused on the classification of voiced and unvoiced phones, keeping the computational cost as low as possible.

References

[1] T. Kohonen, "The "Neural" Phonetic Typewriter *IEEE Computer*", Vol.21, No.3, 1988, pp. 11-22.

[2] M. Malcangi, "Improving Speech Endpoint Detection Using Fuzzy Logic-based Methodologies", in: *Proceedings of the Thirteenth Turkish Symposium on Artificial Intelligence and Neural Networks, Izmir, Turkey, June 10-11, 200.*

[3] M. Malcangi, "Soft-computing Approach to Fit a Speech Recognition System on a Single-chip", in *2002 International Workshop System-on-Chip for Real-Time Applications Proceedings*, Banff, Canada, July 6-7, 2002.

[4] D. O'Shaughnessy, "Speech Communication – Human and Machine", *Addison-Wesley,* Reading, MA, 1987.

[5] C. Hale and C. Nguyen, "Audio Command Recognition Using Fuzzy Logic"*,* Wescon 95, San Francisco, CA, November 7, 1995.

[6] Y. Ying and P. Woo, "Speech Recognition Using Fuzzy Logic"*,* IJCNN '99 – International Joint Conference on Neural Networks, volume 5, 10-16 July 1999, Pages: 2962-2964, vol. 5.

[7] J..Z. Gros, "Text-to-speech synthesis for embedded speech communicators", *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, 2006, pp. 189-193.

[8] Z. Orhan and Z. Gormez, "The framework of the Turkish Syllable-based concatenative text-to-speech system with exceptional case handling", *WSEAS Transactions on Computers*, Vol. 7, No. 10, 2008, pp. 1525-1534.

[9] A. Conkie, "A robust unit selection system for speech synthesis", "Collected Papers of the 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association: Forum Acusticum" (Berlin, Germany, 1999) p. 978; *IEEE transactions on speech and audio processing*, 2001, vol. 9, no. 3, p. 232.

[10] L.R. Rabiner, M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances", The Bell System Technical Journal, Vol. 54, No. 2, 1975.

[11] L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon, "An improved endpoint detector for isolated word recognition", *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-29(4), 777-785, 1981.

[12] Z. Ciota, "Improvement of Speech Processing Using Fuzzy Logic", Approach. In Proceedings of IFSAWorld Congress and 20th NAFIPS International Conference, 2001.

[13] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D., "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy", IEEE International Joint Conferences on Computer Information, and Systems Sciences, and Engineering (CISSE'08).

[14] K. Wang, "A Wavelet-Based Voice Activity Detection Algorithm in Variable-Level Noise Environment", *WSEAS Transactions on Computers*, Vol. 6, No. 8, 2009, pp. 949-955.

[15] D. Impedovo, M. Refice, "Frame Length Selection in Speaker Verification Task", *WSEAS Transactions on Systems*, Vol. 10, No. 7, 2002, pp. 1028-1037.

[16] H. Marvi, "Speech Recognition Through Discriminative Feature Extraction", *WSEAS Transactions on Signal Processing*, Issue 10, Vol. 2, October 2006, pp. 1364-1370.*Systems*, Vol. 10, No. 7, 2002, pp. 1028-1037.

Prof. Mario Malcangi received his undergraduate and graduate degrees in Electronic Engineering and Computer Science from the Politecnico di Milano in 1981. He is member of the International Neural Network Society and among the founders of the Engineering Applications of Neural Networks Special Interest Group (SIG). His research is in the areas of multimedia communications, digital signal processing, and embedded/real-time systems. His research efforts are mainly targeted at speech- and audio-information processing, with special attention to applying soft-computing methodologies (neural networks and fuzzy logic) to speech synthesis, speech recognition, and speaker identification for implementation on deeply embedded systems. He teaches digital signal processing and digital audio processing at the Università degli Studi di Milano. He has published several papers on topics in digital audio and speech processing.