IEEE Instrumentation and Measurement
Technology Conference
Ottawa, Canada, May 19-21, 1997

Speech Recognition Through Phoneme Segmentation and Neural Classification

O. Maeran, V. Piuri, G. Storti Gajani
Dipartimento di Elettronica e Informazione
Politecnico di Milano
piazza L. da Vinci 32, 20133 Milano, Italy
Phone +39-2-2399-3623
Fax +39-2-2399-3411
Email {piuri, storti}@elet.polimi.it

Abstract - The problem of speech recognition may be viewed as the identification of the basic components (phonemes) of the human speech through a high-level measurement procedure working on segments of the vocal signal, their classification, and the identification of their combination into the individual words so as to derive the complete word identification. The efficient and effective solution of the whole problem has several applications (e.g., the automatic typewriter) as well as the simple phoneme recognition may be exploited for some innovative areas (e.g, the voice compression for telecommunication). This paper shows the use of hybrid soft-computing techniques for the signal segmentation and the phoneme classification and the application to voice compression.

I. INTRODUCTION.

Multimedia technologies are becoming widely diffused in the man-machine interaction due to their simplicity and direct understanding for the final user. At the moment, they are adopted only for enhancing the interface capabilities of personal computer (i.e., mainly for office automation, education, publicity, information dissemination, electronic marketing, hobbies, and so on), but in the near future they will be used also to support intensively and extensively other kind of activities such as communication and education for handicapped people, for command delivery, for system control and monitoring, for telecommunication.

Critical aspects in the multimedia technologies are the analysis, the understanding and the storing of huge amount of data, as well as the necessity of treating them in (often hard) real time. Speech treatment and understanding are some of the difficult task in these areas. This paper is specifically focused on the identification of the basic components of the human speech so as to be able to derive information for data transfer, storing and understanding, from the level of the individual component up to the complete words and sentences. In other words, we view the speech recognition as the cascade of the following activities: the identification of the basic components (i.e., phones, represented by phonemes) of the human speech through a high-level measurement procedure working on segments of the vocal signal, their classification, and the identification of their combination into the individual words so as to derive the complete word and sentence identification. In this paper, we focus our attention on the definition of the measurement procedure for a correct identification and classification of the phonemes as well as on the combination of such components to identify the words.

The variability and the complexity of the operations performed by the human brain on the speech to extract such information cannot be easily captured in a traditional completely-algorithmic description, in particular when real-time constraints are imposed for the envisioned application. As a consequence, we experimented a hybrid soft-computing environment to afford each part of the problem with the techniques that have been shown more effective and efficient in the literature for the given application characteristics and constraints. In our opinion, each pure soft-computing techniques (such as the neural networks, the fuzzy logic, the expert systems, and the genetic algorithms) is often not able to deal perfectly with the whole problem, while it may have outstanding abilities for a specific task. Therefore, we mixed algorithmic techniques and neural technologies to achieve signal segmentation, phoneme classification, and data compression, within the application constraints.

Section 2 introduces the characteristics of the speech processing application and the final compression target for telecommunication enhancement and throughput speed-up. Section 3 discusses the overall structure of the measurement and identification procedure, while Section 4 and 5 deal with the specific technological aspects related to segmentation and classification. Section 6 treats the use of our hybrid soft-computing approach to the case of voice compression for telecommunication and data storing.

0-7803-3312-8/97/$5.00©1997 IEEE

II. VOCAL SIGNAL, PHONES AND PHONEMES.

The human speech has a high variety due, for example, to the speaker, the health, the message, the words, the context, the speaking speed, the use of dialects, the environmental noise, the external conditions. In the literature, speech analysis and recognition have been afforded by using several approaches: most of them are based on the identification of the phonemes, i.e., the basic components of the vocal signal. The vowels have been proved in the literature to be quite suited for the automatic detection and recognition since they are usually long enough and have sufficient energy to be identified in the vocal signal. The consonants indeed are much more difficult to be observed in the signal and separated, due to the limited energy and to the number of allophones in the common speech; these problems induced many researchers and industries to discard such an approach.

The origin of the difficulties are actually in the complexity of the separation in phonemes of the vocal signal and in the analysis of the constant-size signal windows.

The human speech is a vocal signal characterized by a spectrum containing several harmonics of the fundamental ones. Position, amplitude, and number of harmonics define the characteristics of the voice. The lungs and the diaphragm produce the air flow. The larynges and, in particular, the vocal cords generate the fundamental harmonics; the vibrating frequency determines the pitch of the voice (50-250 Hz for males, up to 500Hz for females). The oral cavity, the nasal cavity, and the pharynx are the resonant cavities that (through the lips, the teeth, the tongue, and so on) add the higher harmonics and modulate the vocal signal.

The characteristic parameters of the sounds are determined by the involved anatomic structures and by their operation. The number and the interdependencies of these elements induce a high variety and variation of the speech signals and, as a consequence, clearly explain the difficulties occurring in automatic speech recognition. Additional complexity is due to the intrinsic differences in the anatomic structures and in the human control of their operation among different speakers. Besides, the phonetic characteristics of the human speech are strongly dependent on the considered language (in this research, we refer to the Italian language).

The sound generated by phonation is called voiced, while the other sounds are classified as unvoiced. The speech is called connected when it is composed by well-declaimed words which are well separated by short silent intervals. When we speak normally, we often tie adjacent word together and, sometimes, trailing word sounds are eliminated in favor of a better speech fluency; this is called continuous speech. In this paper, we focus our attention on connected speech, as in most of the literature and in commercial tools.

The elementary sounds composing the human speech are called phones. They are the distinct sounds that we are able to recognize. When any isolated phone is pronounced, the shape of the vocal signal is constant for the specific speaker. It changes when the phone is pronounced within the words: our vocal organs cannot in fact change position instantaneously, but need a smooth transition which induces transient shapes in the signal. This is a great difficulty for recognition. An example is shown in Fig. 1 for the letter *ℓ*.

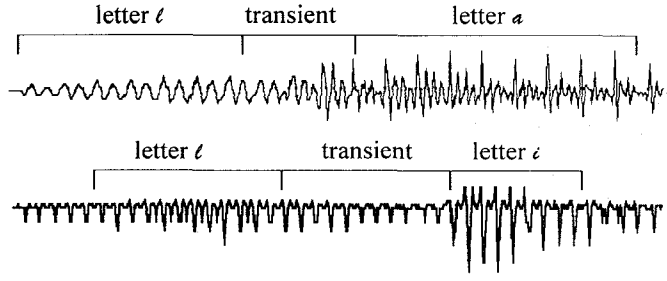

Fig. 1 - Two phones for the same phoneme

To compact the sound classification, we can represent each group of similar phones as a specific phoneme (e.g., all sounds associated to the pronunciation of the letter *ℓ* are represented by the phoneme */l/*). Grouping is performed mainly on the similarity of the sound signals. In our research, we adopted the standard phonetic SAM classification of the Italian language, containing 7 vowels, 2 semivowels, 6 plosives, 4 affricates, 5 fricatives, 3 nasals, and 3 liquids. In our research, we simplified the phoneme classification by removing the unnecessary distinction between the open and closed vowels in order to reduce the recognition complexity without affecting the overall results derived by our approach.

Phones of a phoneme may appear as different sounds: each phoneme has some variants, called allophones. A word remains understandable when an allophone is replaced by another (even if the sound is slightly different), but the word meaning may be changed by replacing a phoneme.

The problem of segmenting the vocal signal and, then, recognizing the basic speech elements is therefore related to the identification of the phonemes. Each phoneme has an its own shape, amplitude, and duration. A clear distinction among these characteristics should allow to separate and recognize the phonemes easily. In general, consonants are the main information carrier, while vowels give energy to the signal for propagation in the air and may carry semantic information. From a phonetic point of view, phonemes can be classified in separate sets, according to the position and the

symbol	phoneme	symbol	phoneme	symbol	phoneme
/d/	/d/+/ʃ/	/m/	/m/	/a/	/a/+/ɟ/
/e/	/e/+/ɛ/	/n/	/n/	/æ/	/S/
/a/	/a/	/ɲ/	/G/	/p/	/p/
/o/	/o/+/O/	/ʃ/	/ʦ/+/ʤ/	/t/	/t/
/u/	/u/+/w/	/ʧ/	/tS/	/c/	/k/
/r/	/r/	/ɟ/	/dZ/	/k/	/k/
/l/	/l/	/ʎ/	/ʎ/	/d/	/d/
/ɟ/	/L/	/v/	/v/	/g/	/g/

Tab. 1 - The classified phoneme for the Italian language

operation of the phonetic organs. In Tab. 1, the phoneme classes considered in our research are given.

III. THE IDENTIFICATION TECHNIQUE.

The phoneme detection and classification have been afforded by using an innovative approach in order to overcome the drawbacks of the traditional techniques, while considering the specific goals of the final application in the areas of speech recognition and speech compression. In particular, the key idea was the adoption of a variable-size adaptable window for vocal signal analysis.

The overall scheme of our approach to speech analysis and phoneme recognition is shown in Fig. 2. The first step of our approach is the acquisition of the voice signal, by sampling it at 8kHz; the digital representation of the voice signal is composed by 8-bit samples, delivered at a 64 kbit/s bit rate. The voice signal is partitioned into segments by using an algorithmic approach based on the variable-size window analysis; the voiced/unvoiced regions are identified as well as the specific phonetic regions within each word and the signal periodicities. The regions are normalized and the specific features of the phonemes are extracted from the segmented signal in order to characterize the individual phoneme; the numeric patterns summarizing the phoneme characteristics are generated by using an algorithmic approach. The subsequent phase is the phoneme classification based on the pattern analysis: it is performed by using a neural approach, due to the difficulties in defining a comprehensive algorithmic method while the separation in classes is quite straightforward from the actual examples.

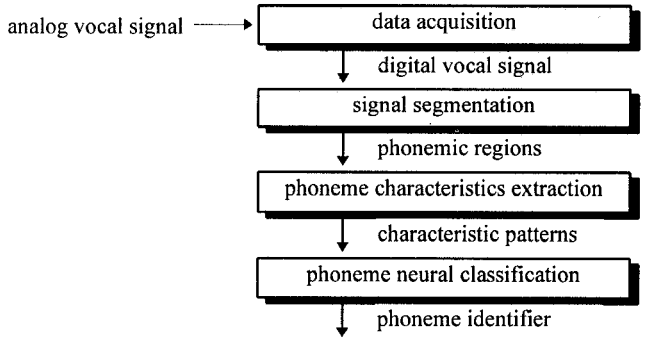

Fig. 2 - The overall algorithm for speech analysis

Finally, we experimented the exploitation of the phoneme recognition and classification for the speech compression problem. Compression, decompression, and voice reconstruction can be realized by means of traditional algorithmic approaches, as we will see in Section 6.

IV. VOICE SEGMENTATION.

We propose an algorithm for the voice segmentation based on the time domain analysis as well as on the use of some ideas for pitch recognition presented in [1]. Once single words have been separated by using the short interval between words, the phoneme regions are extracted in order to be subsequently recognized. When we speak, we produce sounds of many different natures: it is therefore quite difficult to devise a single algorithm capable of segmenting the speech. In our approach, the speech is first segmented in regions that, in some ways, correspond to a division in vowel and consonant sounds; different algorithms are then applied to each type of region for detailed partitioning. The voiced regions require syllable extraction in order to be segmented in phoneme regions, while a different approach for direct phoneme extraction is used for unvoiced regions. The overall algorithm is shown in Fig. 3.

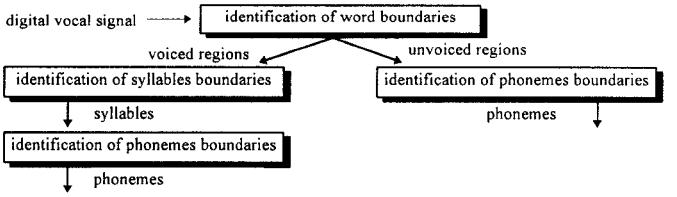

Fig. 3 - The segmentation algorithm

If we analyze a small fragment of the time domain representation of a sampled voice signal (e.g., see Fig. 4), the voiced and unvoiced regions are immediately evident.

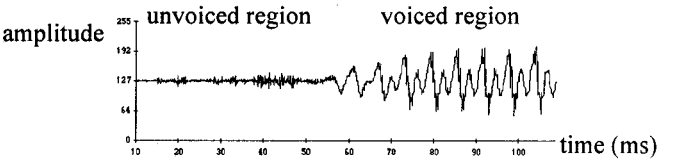

Fig. 4 - The voiced and unvoiced regions

Voiced regions have a much higher energy content and manifest periodicity; on the other hand, the unvoiced regions are more similar to noise. Separation of the different regions is performed in two steps. In the data collection phase, the signal is analyzed to recognize the excursion cycles, i.e., the portion of wave between two zero crossings. Each excursion cycle is examined and the relevant information evaluated and stored; in particular, for each cycle, the following quantities are considered:

- the maximum value;
- the index of the maximum value, both absolute and relative to the first zero crossing;
- the duration, i.e., the number of samples in the cycle;
- the energy, evaluated approximately as the sum of the modulus of each sample;
- the polarity; since cycles may be either positive or negative, it may be defined as the sign of the sum of all samples in a cycle.

The principal cycles are the ones having the maximum energy content within one full pitch period. In Fig. 5, a fragment of the speech signal is shown: four pitch periods

may be easily observed and visually isolated; each period is composed by eight excursion cycles (four positive and four negative), being the four principal cycles marked by an arrow. The segmentation algorithm must recognize the principal cycles automatically.

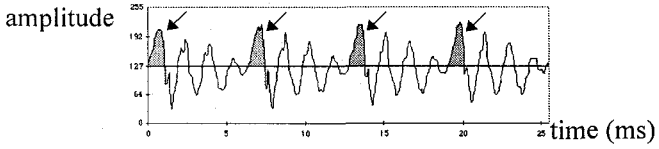

Fig. 5 - Principal cycles

Some characteristics of the voiced and unvoiced regions can be easily inferred by observing speech signals: in the voiced regions, the periods between zero crossings are usually longer, while the cycles have a much higher energy content. This means that the principal cycles can be found only in the voiced region. To recognize the principal cycles, the highest sample value (in modulus) and the highest value of the cycle energy are first identified. The candidates to be principal cycles are all those cycles having an energy content of at least 15% of the maximum energy and an amplitude of at least 20% of the maximum sample value. Due to the nature of the speech, we may analyze only those cycles having the same polarity of the cycle with the maximum amplitude. This first step performs a preliminary screening of the cycles: it is necessary now to select only the principal cycle of each pitch period. Constraints on the period length are introduced: if two cycles are closer than 2 ms, the one with the lower energy content can be discarded; if the interval between two possible principal cycles is greater than 20 ms, such interval represents a discontinuity in the voice signal and contains an unvoiced region. These constraints on the period are derived from the frequency characteristics of human voice: speech rarely displays a pitch frequency smaller than 50Hz or greater than 500Hz. The unvoiced regions are now easily recognized as all the regions without principal cycles. The last step of the segmentation procedure simply eliminates all voiced regions having less than three possible principal cycles and marks them as unvoiced regions.

We define syllable as the section of a voiced region that is centered upon a syllabic nucleus, i.e., a maximal energy principal cycle. This definition, that appears different from the usual one, allows to achieve quite similar results. Each voiced region may contain one or more syllable. To identify them, we must first find all syllable nuclei and, then, recognize the boundaries between syllables. For each voiced region, the first nucleus is easily found as the principal cycle with maximal energy. Other nuclei, if present, must satisfy two conditions:

1. two nuclei must be separated by at least 80 ms,
2. a syllable nucleus has the largest amplitude in a 60 ms region.

By starting from the maximal nucleus and analyzing the signal in both direction, we can easily isolate other nuclei which may be present in a voiced region. This is performed for all voiced regions. The syllabic boundaries are then found by creating the envelope of the speech signal between two nuclei with a 20 ms resolution. The smallest amplitude value is identified: the cycle corresponding to this value is the boundary between two syllables.

Finding the phoneme windows in each syllable is the last segmentation step for the voiced regions. As already said, it is not necessary to extract the exact phonemes, but only the regions with enough information content to allow for the phoneme identification. The differences in the pitch period are not usually sufficient to solve this problem: in many cases, the pitch is identical for a whole voiced region, even if such a region is composed by more than one syllable. The separation criteria that has to be used is more concerned with the shape of each excursion cycle, than with its period. Two criteria are actually used:

1. two principal cycles are similar if the last one differs in amplitude and duration no more than 20% with respect to the first one,
2. two pitch periods are similar if their principal cycles satisfy condition 1 and if their energy differs no more than 15%.

A phoneme window is then defined as the minimal set of periods having similar pitch. The number of periods which are necessary to define this set depends on the average pitch of the speaker; usually, it ranges from 3 to 5. Once windows have been identified, possible oversegmentation problems can be overcome by merging the adjacent windows having similar pitch periods and excursion cycles.

Finding the phoneme windows in the unvoiced regions is somewhat more difficult. First of all, we must remember that these regions may or may not contain phonemes. As already said, the unvoiced regions are identified by two conditions:

1. the duration is greater than 20 ms,
2. no excursion cycle may be considered as principal (excluding the "short" regions that have less than three cycles and that have been marked as unvoiced in the first part of the algorithm).

The first problem to be solved is related to the fact that the transitions from the voiced to the unvoiced regions and vice versa do not usually have a well defined boundary. As shown in Fig. 6, in the voiced regions, we may have some cycles that slowly decay towards the unvoiced part.

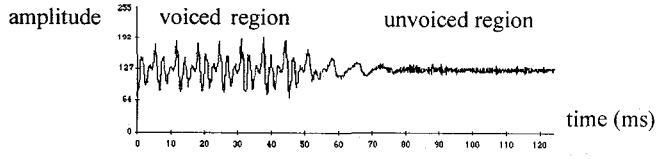

Fig. 6 - Termination of the voiced region

Different criteria must therefore be used: to discard the transition cycles, we adopt a threshold technique similar to the one used to identify the principal cycles. Usually, the transition cycles fall in the unvoiced area due to their low energy content; these cycles are analyzed: if they have at least 20% of the energy of any cycle of any voiced region, they are discarded. A new unvoiced phoneme window starts after the last cycle that has been discarded. If the discarded region lasts more than 20 ms (even if this case is rather rare), it will be considered as a phoneme. A second criteria is given by the silence: the absence of a signal for at least 10 ms bounds phonemes in the unvoiced regions. As in the previous case, the areas bounded by silence must last more than 20 ms to be considered phonemes.

V. THE PHONEME CLASSIFICATION.

To recognize the phonemes, a LVQ neural network has been adopted [2,3]. Some preprocessing has to be done in order to transform the phoneme window in data vectors suitable as inputs for the LVQ network. A first problem is the variable length of each window: usually the phonemes range between 15 ms and 100 ms. To overcome this problem, the signal is normalized by sample decimation or by interpolation so that each window has a constant length. After extensive simulations, this length has been determined equal to 48 ms (384 samples at 8kHz sampling rate). The resulting signal is now framed in order to extract its frequency domain characteristics; 128 samples per frame have given the best results in our simulations. If we overlap each frame by about 50%, five frames covers each normalized window. Each frame is weighted by a Hamming window and the first 15 cepstral coefficient [4] are then extracted. Each phoneme window is represented by a total of 75 cepstral coefficients: they are used as the input patterns to the neural network.
Training of the LVQ network is accomplished by building a 330 element codebook, extracted from an initial training set of 240 words. In the LVQ networks, each codebook vector is equivalent to a neuron in other types of networks. Vectors are initialized randomly. Each word of the training set is segmented in phonemes and preprocessed. The resulting pattern, composed by its 75 cepstral coefficients, is compared to the codebook vectors: the vector having the minimum distance from a particular word is updated. In other words, if we define as $x(t)$ a particular input pattern and as $m_i(t)$ the i-th codebook vector, the nearest vector $m_c(t)$ is defined by:

$$\|x(t) - m_c(t)\| = \arg\min_i \{\|x(t) - m_i(t)\|\}$$

This vector is updated by using the rule:

$$m_c(t+1) = m_c(t) + \alpha(t)\left[x(t) - m_c(t)\right]$$

if $m_c(t)$and $x(t)$belong to the same class (i.e., the code vector and the input pattern correspond to the same phoneme). Otherwise, we use the rule:

$$m_c(t+1) = m_c(t) - \alpha(t)\left[x(t) - m_c(t)\right]$$

This process is repeated for 30000 iterations for each word. The $\alpha(t)$ training coefficient (ranging in 0 to 1) is empirically determined by means of the classical rules of the LVQ networks; a large value is usually adopted during the first iterations, while a small one is preferred for the last ones. We adopted a training algorithm which is different from the standard one described in [3]; our optimized version is based on an adaptive training coefficient and is able to achieve a much faster training. The resulting LVQ network has be tested on 96 words corresponding to 622 phonemes. The results of this test are displayed in Tab. 2.

Phoneme classes	vowels	liquid	nasal	affricate	fricative	plosive
male	90.6	45.8	48.0	56.0	57.9	55.7
female	90.1	43.1	41.3	63.0	24.4	24.1

Tab. 2 - Percentage of correct recognition

The weighted average of the correct phoneme recognition is 75%, which is comparable to the best results available in literature and on the market [5]: such papers present results ranging from 50% to 70%. This result may seem not particularly good from an absolute point of view. However, if an appropriate vocabulary is used, it can be dramatically improved: a correct classification percentage better than 90% has been obtained in preliminary simulations by using a limited vocabulary consisting both of all words in the test set and words having a similar phoneme content. Full intelligibility is anyway obtained even without the vocabulary: our language is in fact highly redundant and a 75% of success in phoneme recognition is more than enough to guarantee the complete speech understanding. Moreover, the unrecognized phonemes are usually identified as other phonemes having anyway a similar sound.

VI. THE APPLICATION TO VOICE COMPRESSION.

An attractive application is voice compression for telecommunication. As already said, the vocal signal can be represented by a sequence of digital samples at a 64 kbit/s rate. In the literature, voice compression is performed traditionally by means of the encoding of the sampled signal. A class of techniques (e.g., DPCM, ADPCM, DM) is based on the waveform encoding: they are able to reduce the bit rate to only few thousands of bits/s. However, such techniques have not been designed specifically for the human voice, but for any sound: as a consequence, they preserve a great amount of information, much more than what is necessary to reconstruct an intelligible voice. Another class of techniques is based on the identification of the voice characteristics parameters so as to avoid all unnecessary information (e.g., Vocoder, LPC, CELP): they are able to

limit the bit rate to only 400 bits/s, even if the quality of the reconstruction may be greatly affected.

In our approach, we consider the voice coding through phonemes, which belongs to the parametric encoding class. Experiments have shown that a speaker is able to generate at most 10-15 phonemes per second due to the time required to change the position and the operation of the vocal organs. Since the Italian language has 30 phonemes, a straightforward coding can be obtained by using 5 bits only for each phoneme. In the worst case, this lead to a maximum bit rate of 75 bits/s: which is definitely smaller than any approach available in the literature and on the market. More efficient encoding scheme can be adopted by using variable-length codes in which shorter codewords are associated to highly frequent phonemes; to such a purpose, we adopted the Huffman coding [6], having an average length equal to 4.065 for the phoneme of the Italian language. With this code, the bit rate is reduced to 61 bit/s.

Therefore, from the input sampled voice, with phoneme encoding, we can achieve a compression ratio of about 1000/1.

Since the coded voice is a bit stream, we can consider a further compression based on the use of traditional digital compression techniques which are typical of data compression in computers and computer networks (e.g., Arc, Arj, BTW, Gzip, LZ, Markov model).

The massive compression based on our phoneme classification technique can be obtained since the phoneme represent the essence of the vocal message. Therefore, transmission of the coded voice on telecommunication networks as well as storing in multimedia system occupies a very limited amount of resources.

However, we have to point out that this compression removes all redundant information that could be semantically meaningful (e.g., the voice inflection, the accents, the speed, the speaker identification). This is a great drawback when the voice need to be reconstructed. A partial possible solution consists of correcting the voice reconstruction by using the specific phonemes sampled on the voice of the speaker, instead of a generic set of phonemes. The characteristics phonemes of the speaker can be sampled and stored in a correction table; this information can be transmitted to the computer system of the listener before the speaker begins (or retrieved from a file on such a system). The resulting sound is still far from the original uncoded voice since it is discontinuous and allophone can affect the intelligibility of the reconstructed voice. This problem can be dealt with by introducing additional information in the phoneme table, e.g., about the transitions between the phonemes (called diphones) being such information very critical in the reconstruction phase; for the Italian language, it can be seen that only 150 diphones are necessary.

VII. CONCLUSIONS.

A new technique for speech analysis and processing has been presented to deal with the identification and classification of phonemic regions in the connected human speech, by using variable-size windowing and neural networks. The results achieved have suggested that our approach is quite effective to solve this problem: our phoneme recognition leads to a word identification which is comparable to the techniques available in the literature and on the market, even if we do not add any post-processing technique to refine the classification by taking into account additional context-dependent information which are typical of the speech redundancy and human speech understanding. A great enhancement of overall speech recognition could in fact be obtained by adding, e.g., orthographic and grammatical correctors; this will lead to better performances than the one currently available. These results are partially a point against the current opinion that the phoneme-based recognition is not a feasible solution, even if it is not easy, similarly to other statistical approaches. Besides, we have to point out that the phoneme-based recognition is much faster than such other techniques, so that it is much more suitable for real-time applications.

On the other hand, the phoneme recognition ability of our approach has been shown effective to solve the voice compression problem by itself, while the addition of further traditional compression techniques to the binary representation of the coded phonemes may allow to further increase the compression ratio. Some problem concerning the quality of the decompressed voice are still open: interpolation and transition techniques are presently studied to smooth the cascading of subsequent phonemes into a connected reconstructed speech.

REFERENCES

[1] N.J. Miller, "Pitch detection by data reduction", *IEEE Trans. On ASSP*, vol. 23, n. 1, Feb. 1975.

[2] T. Kohonen, *Self-Organization and Associative Memory*, Spriger Verlag, Berlin, 1989.

[3] T. Kohonen, et al., "LVQ_PAK. The learning vector quantization program package", otaniemi ftp site, version 3.1, 1995.

[4] A.V.Oppenheim, R.W. Shafer, *Digital Signal Processing*, Prentice Hall, 1975.

[5] R. Linggard, D.J. Myers, C. Nightingale, *Neural networks for Vision, Speech, and Natural Language*, Chapman & Hall, 1992.

[6] R.W. Hamming, *Coding and Information Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1980.