Proceedings of the Second Conference on Speech Technology and Human-Computer Dialogue (SpeD 2003)
Bucharest, Romania, April 10-11, 2003 (accepted for publication)

ROBUST RECOGNITION OF SMALL-VOCABULARY TELEPHONE-QUALITY SPEECH

Dragos BURILEANU[*], Mihai SIMA[**], Cristian NEGRESCU[*], Victor CROITORU[*]

[*] University "Politehnica" of Bucharest, Faculty of Electronics and Telecommunications
[**] Delft University of Technology, Department of Electrical Engineering

Corresponding author: Dragos BURILEANU bdragos@mESsnet.pub.ro

Considerable progress has been made in the field of automatic speech recognition in recent years, especially for high-quality (full bandwidth and noise-free) speech. However, good recognition accuracy is difficult to achieve when the incoming speech is passed through a telephone channel. At the same time, the task of speech recognition over telephone lines is growing in importance, as the number of applications of spoken language processing involving telephone speech increases every day. The paper presents our recent work on developing a robust speaker-independent isolated-spoken word recognition system based on a hybrid approach (classic – artificial neural network). A number of experiments are described and compared in order to evaluate different analysis and recognition techniques that are best suited for a telephone-speech recognition task. In particular, we address the use of RASTA processing (i.e., filtering the temporal trajectories of speech parameters) for increasing the recognition accuracy. Also, we propose a method based on the adaptive filter theory for producing simulated telephone data starting from clean speech databases.

Key words: robust speech recognition, telephone-quality speech, Learning Vector Quantization, adaptive filtering, RASTA processing

1. INTRODUCTION

Important efforts are being carried out to continuously improve the performance of speech recognition systems. A distinct research activity is focused on recognition in adverse conditions, that is, when the quality of the speech at testing phase is lower than speech at training phase. This area is generally referred to as *robust speech recognition*, where robustness refers to the needs of maintaining good recognition accuracy even when the quality of the input speech is degraded [6], [10], [11], [22].

This paper presents our current research and achievements on recognition of small-vocabulary speaker-independent isolated-spoken words over the telephone line. We first outline the basic architecture of our speech recognition system based on a classic – artificial neural network (ANN) hybrid approach, and provide recognition performance when this system has been trained and tested with clean speech. Since a professional acoustical database recorded over the telephone line is not available for Romanian language, we propose to simulate the telephone channel by means of an adaptive filter and to emulate a telephone-quality acoustical database starting from clean database. The main idea is to determine the coefficients of a FIR filter starting from two available professional acoustical databases: TIMIT (clean speech) and NTIMIT (a version of TIMIT recorded via telephone line). We also address the transformations that are encountered by a cepstral-based representation when speech is transmitted over the telephone line, and discuss the use of RASTA processing [8] for channel normalization.

The paper is organized as follows. We discuss several topics related to isolated-spoken word recognition without time alignment, and present the overall organization of our recognition system along with main implementation issues in Section 2. The strategy of simulating the telephone channel is described in Section 3. The experimental framework and results are presented in Section 4. Section 5 completes the paper with some conclusions and closing remarks.

2. BACKGROUND

To make the presentation self-consistent, we would like to address some issues related with isolated-spoken word recognition, highlighting those recognition techniques which alleviate temporal alignment. We will also review the architecture of our speaker-independent isolated-spoken word recognition system, emphasizing the parametrical representation of speech that led to the highest recognition rate with respect to the mentioned architecture.

2.1. Isolated-spoken word recognition without time alignment

The recognition of isolated-spoken words is theoretically less difficult than continuous or spontaneous speech recognition. Moreover, one can use *global recognition techniques* having the word as the basic unit and therefore simplifying the task of the recognition system, especially for small and medium vocabularies.

However, isolated word recognition is not a trivial task, at least in two classes of applications: *speaker-independent recognition*, or / and *robust isolated word speech recognition*. If the latter topic will be discussed afterward, about the former one we would like to mention that one need to find an optimum feature set that can: (1) characterize the words from the vocabulary as accurate as possible (even in adverse condition), and (2) handle a large amount of labeled (training) speech data in order to track the large speaker and speech variability. Specifically, important variations in speaking durations and rate, for different speakers, or even for the same speaker at different moments frequently occur.

A classic method used in the classification stage to compensate these important temporal variations in speech utterances is to compute a local distance between two parameter vectors simultaneously with a *global time alignment* score. This global alignment score is computed by accumulating the local distance such that an overall distortion is minimized. This concept, which is known as *dynamic time warping* (DTW) [15], is relatively simple to implement. However, since it is strictly a sequential algorithm, it does not exhibit *instruction level parallelism* [14]; thus, it cannot benefit from implementation support provided by modern massively parallel computing engines. In addition, it is well known that this algorithm is based on a number of local and global constraints (usually rather simple, for an ease of implementation), and therefore it is difficult to find an optimal path for the temporal alignment.

A way to avoid the shortcomings associated to DTW is to use a pattern recognition method based on *vector quantization* (VQ). Vector quantization is an effective method to deal with speaker dependency in the feature space. By performing unsupervised learning, reference vectors that optimally represent the corresponding classes are generated; for small vocabulary recognition systems, training data can be effectively arranged to generate an optimal set of reference vectors for each word class [2], [15].

It is beyond the scope of the paper to elaborate on the whole theory of vector quantization and vector quantizer design. Instead, the scope here is restricted to the issues that seem important for the application of vector quantizers as speech recognition engines.

A vector quantizer divides the parameter space of a multidimensional feature vector into a fixed number of regions, which are referred to as *clusters*, or *partitions*. After a *centroid* computation procedure is completed, each cell is associated with its centroid, also named *reference vector* (or *codeword*); these vectors are organized into a *codebook*. Assuming that $\mathbf{x}(t)$ is a feature vector that defines a short time spectrum at regularly spaced intervals in time and $\mathbb{C} = \{\mathbf{y}_i\}_{i=1}^{M}$ is a set of reference vectors, the quantizer encodes $\mathbf{x}(t)$ by the index i of the reference vector $\mathbf{y}_i$ that minimizes a prescribed distortion measure [15].

The procedure of finding an optimum codebook (or a set of codebooks) with respect to a distortion measure has to be done on a training database in an unsupervised fashion, since the distribution of the feature vectors is unknown. An example of a training procedure is the Linde-Buzo-Gray (LBG) algorithm [13]. Basically, this algorithm is as follows. Starting with a codebook of size 1, the procedure successively doubles the codebook size until the desired number of codebook vectors is obtained. For each codebook size, the codebooks are optimized with *k-means* algorithm until an optimality criterion is met. Then, the number of codebook vectors is doubled by adding a small random vector to each of the existing centroids.

As mentioned, the concept of vector quantization can be easily applied, for example, to isolated-spoken word recognition. Suppose there are V utterance classes (i.e. words) to be recognized. The training vectors which belong to the same class k are used to design a class-oriented vector quantizer. Thus, there are V vector quantizers. During recognition, an unknown spoken word is vector quantized frame-by-frame by all V quantizers, resulting in V sums of distortion scores. The word is recognized as belonging to class k if the minimum sum has been obtained for class k. This is indeed distinct from the DTW algorithm mentioned above.

The vector quantization technique has been used as the basic step in the classification stage for the speaker-independent isolated-spoken word recognition system recently developed in our laboratory (both in the training and recognition phases).

2.2. A speaker-independent isolated-spoken word recognition system

We would like to emphasize that the codebooks generated by a VQ procedure presented in Subsection 2.1 are optimized to introduce the lowest possible distortion to the original vector sequence. If the codebooks are used for classification, however, the main concern is not about low distortion but rather that the partition borders follow as accurately as possible the Bayes decision boundaries between classes. Therefore, using a specific distortion criterion, the above-described procedure could lead to an optimal codebook only for each training set, that is utterances of the same word, but doesn't offer supplementary information about the "class" of that word (i.e. the distinct features of the word in the framework of the given vocabulary) and the classification performance for each word in the vocabulary after obtaining the lowest distortions to each codebook.

As a result, a (classic) VQ-based recognition engine can achieve a high recognition rate only for vocabularies of highly distinct words. For a vocabulary that contains words subject to generate confusion even for a human listener (e.g., voice dialer in Romanian), such recognition method generally does not provide satisfactory recognition performance.

For this reason, we have proposed a neural network approach to further increase the recognition performance, by using a *learning vector quantization* (LVQ) algorithm for the codebook optimization through a supervised procedure, using all the words from the training database [18].

LVQ is one of the best connectionist techniques for classification tasks. Its performances are comparable, for example, to multi-layer networks, but for much reduced learning time [4]. LVQ works as follows: each class is characterized by a fixed set of reference vectors, of the same dimensions as the data to be classified. When an unknown vector is presented, all the reference vectors are investigated to determine the nearest one, in the Euclidean distance sense. The vector is then classified into the class of this nearest reference. LVQ is therefore an algorithm for adaptively modifying the references; in other words, it modify the centroids of a given codebook, using labeled training vectors, so that the partition borders approximate the Bayesian decision borders.

We used the so-called LVQ3 algorithm [12]. The philosophy of the algorithm is simple and elegant. Briefly, it updates a given codebook for the labeled training vector as follows: move the codebook vector with the same class as the input training vector towards the input training vector, otherwise move the codebook vector away from the input training vector.

The organization of our recognition system highlighting the training and recognition phases is presented in Figure 1.

First, the acquisition is carried out using a sampling frequency of 8 kHz (16 bit/sample, linear quantization law), the resulting utterances being stored as "wav" files. The recording time for each utterance is 1 second. During the preprocessing stage, a *start-end point detection* is performed in order to discard the silence regions [17]. In order to compensate for the spectral tilt in the signal spectrum which is due to the shape of the excitation function produced by the glottis, and also to reduce the error when performing a LPC analysis with finite precision, a pre-emphasizing algorithm is carried out with a first-order FIR filter having the transfer function: $H(z) = 1 - 0.95z^{-1}$.

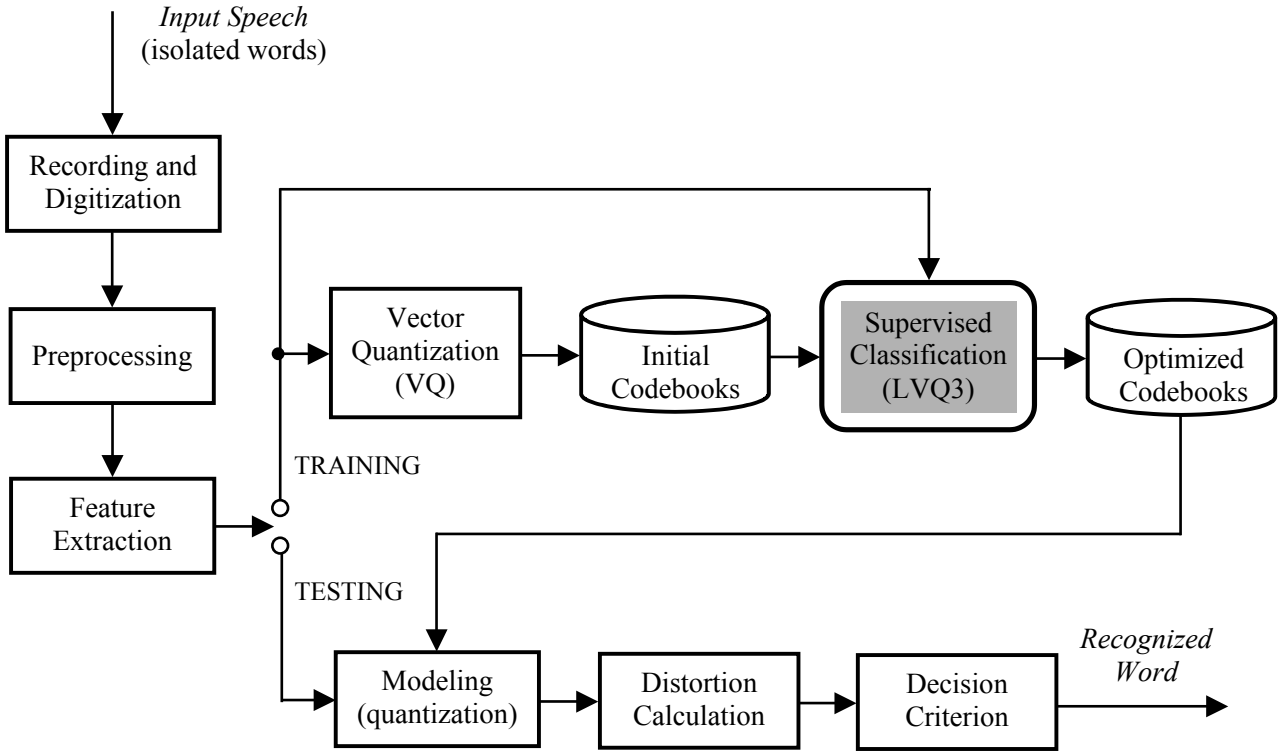

Figure 1. The organization of the proposed speech recognition system (training and recognition phases)

For the feature extraction, the analysis is carried out on short speech segments. We used a frame length of 10 ms (80 samples), with a frame overlap of 5 ms (40 samples). To reduce the effects of the abrupt discontinuities at the edges of the analysis interval, we used a Hamming window having the length equal to the frame length (i.e., 80 samples). A number of different parametric representations have been considered:

- *MFCC*: 12 cepstral coefficients derived from Fourier transform and evaluated on *mel* scale;
- *LPCC*: 12 cepstral coefficients derived from linear prediction spectrum;
- *NLPCC*: 12 *LPCC* coefficients normalized by means of a filter in the cepstral domain;
- Δ*LPCC*: first-order derivatives of *LPCC* coefficients;
- Δ*NLPCC*: first-order derivatives of *NLPCC* coefficients;
- log *E*: logarithm of the energy;
- Δlog *E*: first-order derivative of log *E*.

To train the system, the LBG algorithm has been used to initialize the codebook distribution. Then, LVQ3 algorithm has adjusted the partition borders so that they approximate the Bayes decision borders much better. For a vocabulary containing the ten digits (from 'zero' to 'nouă') in Romanian language and using clean speech databases both for training and testing, the best recognition rate of 98.33% (for the testing database) has been achieved for the combination of 26 coefficients (*NLPCC* + Δ*NLPCC* + log *E* + Δlog *E*) and a codebook size equal to 32. A few more comments about these results will be presented in Section 4. For more details, we refer the reader to bibliography [18].

As mentioned in the Introduction, the concept of robust speech recognition generally refers to maintaining good recognition accuracy even when the quality of the incoming speech is degraded by noise. An example of such degradation is provided by on-line applications, when the speech is transmitted over the telephone line before the recognition task is initiated. In the sequel, we will address this subject and will provide recognition rate figures for a voice dialer in Romanian.

3. SIMULATING TELEPHONE-QUALITY SPEECH

Since a professional acoustical database recorded over the telephone line is not available for Romanian language, we propose to simulate the telephone channel and to emulate a telephone-quality acoustical database starting from clean database. Such approach, which is also used by other researchers [11], [21], allows for a high flexibility of the undergoing tests, as well as SNR variations that are difficult to be provided for a particular acquisition of the acoustical database.

Usually, a simple low-pass filtering and (sometimes) noise addition are used to simulate the speech transmitted over the telephone line. Since the telephone line introduces complex changes in the signal spectrum that cannot be efficiently modeled by a simple low-pass filtering process, we designed a 64-order linear FIR filter that approximates the average response of the telephone channel. For this purpose, we used pairs of signals (spoken utterances) from two professional databases: TIMIT and NTIMIT. The TIMIT speech corpus is a popular database, widely used by the speech research community. It contains utterances collected from a large number of speakers, recorded using a close-talking noise-canceling microphone; recordings are relatively clean with respect to non-speech noise. The NTIMIT corpus was developed to provide a telephone band-limited speech adjunct to the TIMIT database and was collected by transmitting all original TIMIT utterances through various channels in a telephone network and re-digitizing them.

Subsequently, we will propose a strategy to determine a model for the telephone channel. For the sake of presentation, let us denote the original (TIMIT) signal (applied to the input of the channel) with $x(n)$, and the corresponding output (NTIMIT) signal (delivered by the channel) by $d(n)$. Both signals (originally sampled at 16 kHz, 16 bit/sample, linear quantization law) have been firstly down sampled at 8 kHz. Then, we have modeled the channel as a FIR transversal filter having the impulse response:

$$h'(n) = \sum_{i=0}^{M'} w_i' \delta(n-i). \tag{1}$$

The weight coefficients w_i' are computed via an adaptive procedure using the adaptive identification configuration presented in Figure 2.

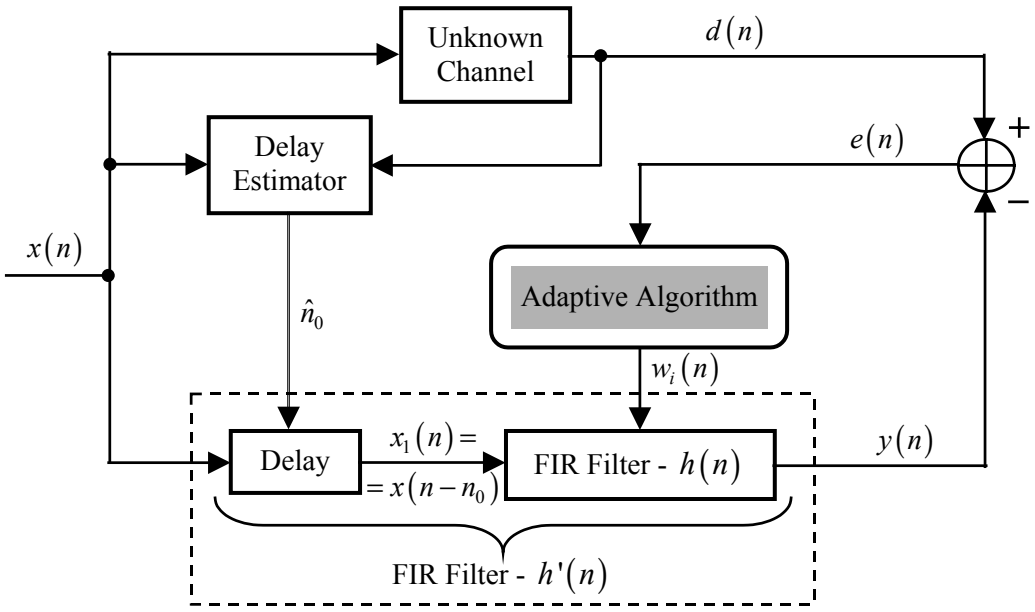

Figure 2. The adaptive identification of the telephone channel

The order M' of the FIR filter should be large enough to allow an accurate modeling of the unknown channel with respect to a prescribed cost function. On another hand, in order to decrease the computation demands and also to improve the behavior of the adaptive algorithm (convergence speed, misadjustment, etc.), M' should be as small as possible [7]. A reasonable compromise has been obtained by taking into account the inherent delay introduced by the channel which yields to zero values of first w_i' coefficients. We split the FIR model of the unknown channel in two parts, therefore we cascaded a pure delay (with n_0 steps) with a FIR filter of order $M = M' - n_0$. The impulse response of this last FIR filter is

$$h(n) = \sum_{i=0}^{M} w_i \, \delta(n-i), \text{ with } M = M' - n_0 \text{ and } w_i = w'_{i+n_0}. \tag{2}$$

The delay n_0 needs to be estimated before performing the adaptive algorithm, which computes the w_i coefficients. In our experiments this operation is based on the short-term cross-correlation between x and d signals [19]. For an accurate estimation, we have selected an unvoiced phoneme with a good time localization (such as 'p', 't', 's') that follows up a silence interval. The estimate of the time delay group expressed in number of samples ($\hat{n}_0$) is given by the localization of the maximum from the previously computed cross-correlation. To include the "non causal queue" we have chosen the actual delay n_0 smaller than $\hat{n}_0$ (in our experiments $\hat{n}_0$ was about 60 samples and n_0 was set to $\hat{n}_0 - 30$ samples).

From now on the computing of the w_i coefficients falls into a classical *adaptive system identification procedure* [7]. We used the *mean square error* (MSE) as an objective cost function (with the physical meaning of the statistical power of the error signal), while for the adaptive algorithm we used the well known *normalized least mean squares* (nLMS) [7]. To implement the nLMS algorithm, we have considered $M = 256$ and the step size $\alpha = 0.05$. The convergence of the algorithm and the assumption for stationary unknown channel are verified using the $M_T SE$ sequence [5]. For each epoch (i.e., each pair of TIMIT/NTIMIT signals), the final w_i coefficients of the impulse response (and, therefore, the final values for w_i') have been obtained by computing the temporal average after the adaptation process completed:

$$w_i = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} w_i(n), \text{ where } n_2 - n_1 = 35{,}000 \text{ and } n_1 > 45{,}000 \tag{3}$$

We performed 32 numerical experiments with different signals pairs from TIMIT/NTIMIT databases. For each experiment, we computed the channel's transfer function by applying a 256-points FFT routine on its impulse response sequence. Due to the perceptual relevance, the magnitude of the transfer function was averaged over the entire number of experiments. Finally, using the theory and procedure presented in [20], a causal and truncated replica (64 samples) of the channel impulse response has been obtained. The frequency response of the 64-order FIR filter is depicted in Figure 3.

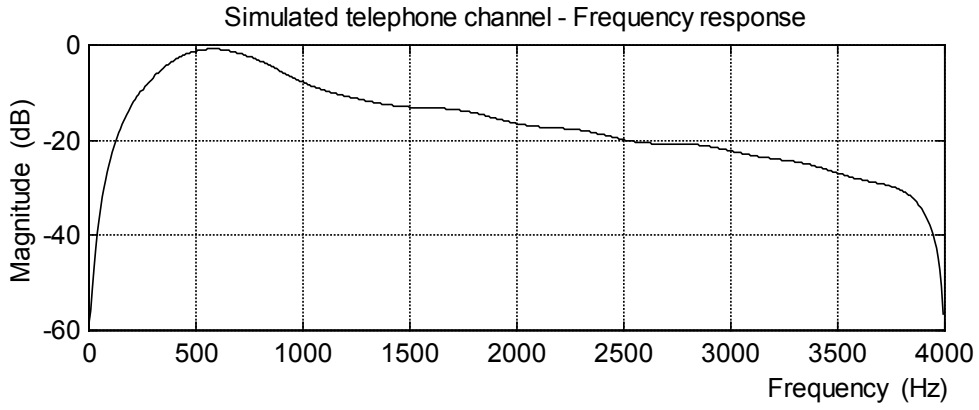

Figure 3. The frequency response of a FIR filter that simulates the telephone channel

4. EXPERIMENTAL RESULTS

The recognition of the Romanian digits is considered as a case study. A *multiple-speaker acoustical database* (MSADB) containing 1,800 words has been recorded in normal laboratory conditions: 20 male speakers and 10 female speakers uttered 6 times each of the ten digits, with two different speaking rates and three distinct prosodic contours. This database has been partitioned in two sub-databases: the *training database*, which includes 1,200 words spoken by 20 speakers (13 men and 7 women), and the *testing database*: 600 words spoken by 10 speakers (7 men and 3 women). More details about the entire structure of the database can be found in [4] and [18].

In addition, we mention that the filter presented in the previous section has been used to filter the testing database (600 utterances). Thus, a database that emulates speech transmitted over the telephone line has been obtained.

The system described in Section 2 has been trained with the training database (1,200 words). Then, a lot of tests have been carried out with the clean testing database and the telephone-quality testing database, in order to find the best parameter set and recognition strategy.

First, to resume the results stated in Subsection 2.2, we present in Table 1 the *word recognition rate* (WRR) obtained for different combinations of parameters, when using the clean testing database.

Table 1. Word recognition rate (WRR) for different sets of parameters (clean testing database)

Parameter set	**WRR [%]**	
	Training database	**Testing database** (clean speech)
MFCC	98.67	93.67
LPCC	98.92	94.33
NLPCC	99.42	95.50
LPCC + Δ*LPCC*	99.75	96.83
NLPCC + Δ*NLPCC*	99.92	97.67
NLPCC + Δ*NLPCC* + log *E* + Δlog *E*	99.92	98.33

Regarding the recognition over the telephone line, witch is the main subject of this paper, we would like to mention that since our choice is for a cepstral-based parametrical representation of speech, we are mostly interested how are the cepstral parameters affected when speech is transmitted over the telephone line. As described in [16], the major effects consist of: (1) shifting of the mean of each cepstral coefficient, (2) reducing of their standard deviation, and (3) exhibiting non-Gaussian distribution. To minimize such errors, two methods have been proposed up to now: *intraframe processing*, a typical example being the *band-pass lifter*, which we used in generating the *NLPCC* coefficients, and *interframe processing*, which is best represented by the *relative spectral processing* (RASTA) [1], [3], [8].

After the telephone-quality testing database has been obtained, we used in the first major test the optimum set of parameters described in Subsection 2.2 (26 parameters) and a codebook dimension of 32. As it is shown in Table 1, for the clean testing database the recognition rate is 98.33 %. Using the simulated database, the recognition rate drops to 94.17 %. This recognition rate is better than those claimed in the literature, not necessarily in absolute value but relative to the rate obtained by testing with a clean database (the degradation of the recognition performance due to the use of a filtered database usually ranges from 5 to 10 times, while our results show a degradation of only 3.5 times).

In the second major test we used RASTA. Theoretically, this processing technique has a large potential due to the both underlining principle (filtering of the slow variations of the cepstral coefficients in order to compensate for the telephone channel variations) and implementation simplicity [8], [9]. We first calculated all the cepstral coefficients on frames of 10 ms for each and every utterance from the (filtered) testing database. The resulted slow-varying signals corresponding to each cepstral coefficients are sent to a band-pass filter having the transfer function:

$$H(z)=\frac{z^4\left(0.2+0.1z^{-1}\right)-0.1z^{-3}-0.2z^{-4}}{1-0.98z^{-1}}. \tag{4}$$

The frequency response and the impulse response of the designed RASTA filter are presented in Figure 4 (the bandwidth of this filter approximately ranges between 1 and 13 Hz).

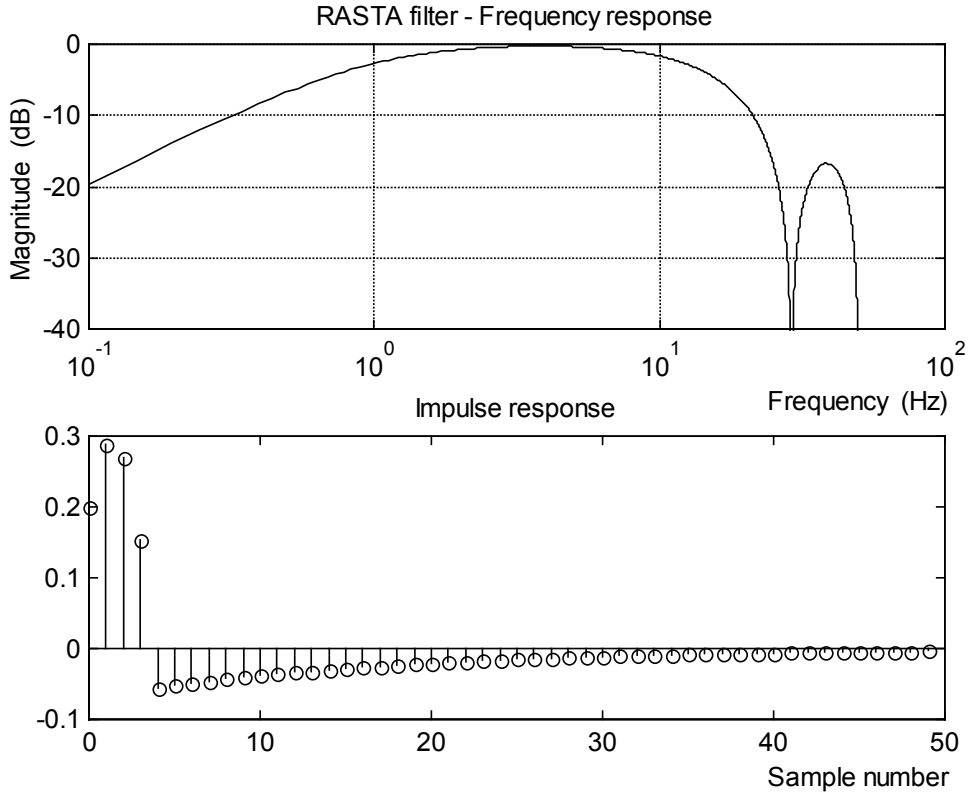

Figure 4. Frequency response and impulse response of the RASTA filter

These new (filtered) cepstral coefficients were afterward used on each analysis frame by the recognition system the same way as the original ones.

The efficiency of the RASTA filtering has been evaluated with different sets of parameters. We determined that the maximum recognition rate is obtained with RASTA-filtered *NLPCC* coefficients (12 coefficients), along with log E + Δlog E (a total of 14 coefficients). The recognition rate is 95.33 %, which is higher than 94.17 % that has been obtained with the 26 coefficients in the conditions described above.

We expect that using RASTA, the relative improvement of the recognition rate with respect to the clean-speech experiment will be similar when a testing database recorded over a real telephone line is used. This can be explained by the fact that RASTA has proved its efficiency for a time-varying telephone channel; even if in our experiment the simulated telephone channel does not have temporal variations, we still used RASTA for compensation of the noise introduced by the acquisition devices and of the cepstral parameter variations due to the (inherent) speaker and speech variability during the recording sessions.

Finally, a third major test has been accomplished in order to evaluate the performance improvement of our system when using (in the training phase) the supplementary codebook optimization through the supervised LVQ3 algorithm. We used the optimum set of 14 coefficients previously mentioned and a codebook size equal with 32. Table 2 shows these final results obtained with, and without codebook optimization, and Table 3 presents the *confusion matrix* (containing the number of errors and their locations) for whole testing database (10×60 words) and the best recognition rate of 95.33%.

Table 2. Word recognition rate (WRR) for two classification strategies (telephone-quality testing database)

Classification strategy	**WRR [%]**	
	Training database	**Testing database** (telephone-quality speech)
VQ	96.08	91.33
VQ + LVQ3	98.66	95.33

Table 3. Confusion matrix for the global word recognition rate of 95.33% (telephone-quality testing database)

Uttered Word	**Decision** (recognized word or writted message)											**WRR [%]**
	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	'Nu_înțeleg!' (*I cannot understand!*)	
'0'	59	0	0	0	**1**	0	0	0	0	0	**0**	98.33
'1'	0	58	0	0	0	0	0	0	0	0	**2**	96.66
'2'	0	0	59	0	0	0	0	0	**1**	0	**0**	98.33
'3'	0	0	0	58	0	0	0	0	0	0	**2**	96.66
'4'	0	0	0	0	59	0	0	0	0	0	**1**	98.33
'5'	0	0	0	0	0	58	0	0	0	0	**2**	96.66
'6'	0	0	0	0	0	0	52	**6**	0	0	**2**	86.66
'7'	0	0	0	0	0	0	**5**	54	0	0	**1**	90.00
'8'	0	0	**1**	0	0	0	0	0	57	0	**2**	95.00
'9'	0	**1**	0	0	0	0	0	0	0	58	**1**	96.66

5. CONCLUSIONS AND FUTURE WORK

The paper discussed the topic of robust speech recognition of small-vocabulary over the telephone line. We first described the architecture and the classification strategy for a speaker-independent isolated-spoken word recognition system. We proposed a method based on the adaptive filter theory to approximate the average response of the telephone channel and used the resulted FIR filter to create a (simulated) telephone-quality speech database. We also analyzed the RASTA processing technique and have use it to filter the slow variations of the cepstral coefficients.

The recognition of the Romanian digits has been considered as a case study. For this particular vocabulary, recognition rate figures have been determined for several parametric representations of speech. The experimental results indicate that the parametric representation that consists of 12 *NLPCC* and 12 Δ*NLPCC* coefficients along with log *E* and Δlog *E* drops the recognition rate from 98.33% (achieved with clean speech) to 94.17% (achieved with telephone-quality speech). At the same time, the RASTA filtering in connection to 12 *NLPCC* coefficients and log *E* + Δlog *E* provides a recognition rate of 95.33%. These results clearly indicate that the later parametric representation along with the architecture of our recognition system and the proposed classification strategy (VQ + LVQ) is a promising approach with respect to small-vocabulary speaker-independent isolated-spoken word recognition over the telephone line.

We would like to mention that although we selected a vocabulary of small size (isolated-spoken utterances of the 0 ÷ 9 digits), the proposed recognition strategy does not depend on the word set and vocabulary size. Thus, in future work we aim to extend the vocabulary size with a few new words suitable for a real voice-dialer application; the only foreseeable constraint is the need for real-time response.

We also intend to generate a supplementary "noise" codebook (including a number of words from outside the vocabulary and nonsense words, along with several types of noise), which will allow the system to accept words that are not part of the designated vocabulary as inputs. This way, the system will become more appropriate for a real-world application.

REFERENCES

1. AVENDANO, C., VUUREN, S.V., HERMANSKY, H., *Data Based Filter Design for RASTA-like Channel Normalization in ASR*. Proceedings of ICSLP'96, Philadelphia, USA, pp. 2087-2090, 1996.
2. BOITE, R., BOURLARD, H., DUTOIT, T., HANCQ, J., LEICH, H., *Traitement de la parole*, Presses Polytechniques et Universitaires Romandes, Lausanne, 2000.
3. BOURLARD, H.A., HERMANSKY, H., MORGAN, N., *Towards Increasing Speech Recognition Rates*. Speech Communication, No. 18, pp. 205-231, 1996.
4. BURILEANU, D., SIMA, M., BURILEANU, C., CROITORU, V., *A Neural Network-Based Speaker-Independent System for Word Recognition in Romanian Language*. Proceedings of The First Workshop on Text, Speech and Dialogue, Brno, Czech Republic, pp. 177-182, 1998.
5. CIOCHINA, S., NEGRESCU, C., *Adaptive Systems***,** Bucharest, Ed. Tehnică, 1999 (in Romanian).
6. GONG, Y, *Speech Recognition in Noisy Environments: A Survey*. IEEE Transactions on Speech and Audio Processing, **Vol. 8**, No. 6, pp. 664-675, Nov. 2000.
7. HAYKIN, S., *Adaptive Filter Theory***,** Fourth Edition, Prentice Hall - Information and System Science Series, 2001**.**
8. HERMANSKY, H., MORGAN, N., *RASTA Processing of Speech*. IEEE Transactions on Speech and Audio Processing, **Vol. 2**, pp. 575-589, Oct. 1994.
9. HERMANSKY, H., *Mel Cepstrum, Deltas, Double-Deltas,... What Else is New?*. Proceedings of Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, **Vol. 1**, pp. 866-871, 1999.
10. HUANG, X., ACERO, A., HON H.W., *Spoken Language Processing. A Guide to Theory, Algorithm, and System Development*, New Jersey, Prentice-Hall, 2001.
11. KARNJANADECHA, M., ZAHORIAN, S.T., *Signal Modeling for High-Performance Robust Isolated Word Recognition*. IEEE Transactions on Speech and Audio Processing, **Vol. 9**, No. 6, pp. 647-654, Sep.2001.
12. KOHONEN, T., *The Self-Organizing Map*. Proceedings of the IEEE, **Vol. 78**, No. 9, pp. 1464-1480, Sep.1990.
13. LINDE, Y., BUZO, L.R, GRAY, R.M., *An Algorithm for Vector Quantizer design*. IEEE Transactions on Communication, **Vol. 1**, No. 28, pp. 84-95, Jan. 1980.
14. PATTERSON, D.A., HENNESSY, J.L., *Computer Architecture. A Quantitative Approach*, Second Edition, Morgan Kaufmann, San Francisco, California, 1996.
15. RABINER, L.R., JUANG, B.H., *Fundamentals of Speech Recognition*, New Jersey, Prentice-Hall, 1993.
16. SHIEH, W.C., CHANG, S.C., *The Dependence of Feature Vectors under Adverse Noise*, Proceedings of EUROSPEECH'99, Budapest, Hungary, **Vol. 5**, pp. 2395-2398, 1999.
17. SIMA, M., BURILEANU, D., BURILEANU, C., CROITORU, V., *Full-Custom Software for Start/End Point Detection of Isolated-Spoken Words*. Proceedings of the 12th International Conference on Control System and Computer Science, Bucharest, **Vol. II**, pp. 19-24, 1999.
18. SIMA, M., *On Neural Networks Utilization for Signal Processing in Telecommunications – Speaker-Independent Recognition of Isolated-Spoken Words in Romanian*, PhD Thesis, University "Politehnica" of Bucharest (in Romanian).
19. STANOMIR, D., NEGRESCU, C., JALBA, L., *Algorithms For Speech Processing – Theory and Applications in GSM Communications*, Bucharest, Ed. Athena, 2003 (in Romanian).
20. STANOMIR, D., NEGRESCU, C., *The Hilbert Transform – From Pure Theory To Practical Applications*. SpeD 2003, Bucharest, 2003 (to appear in this volume).
21. TARCISIO, C., DANIELE, F., ROBERTO, G., *Use of Simulated Data for Robust Telephone Speech Recognition*. Proceedings of EUROSPEECH'99, Budapest, Hungary, **Vol. 6**, pp. 2825-2828, 1999.
22. VETH, J., BOVES, L., *Comparison of Channel Normalization Techniques for Automatic Speech Recognition over the Phone*. Proceedings of ICSPAT'96, Boston, USA, **Vol. 2**, pp. 1784-1788, 1996.