Recognizing Voice Over IP: A Robust Front-End for Speech Recognition on the World Wide Web

Carmen Peláez-Moreno, Ascensión Gallardo-Antolín, and Fernando Díaz-de-María, *Member, IEEE*

***Abstract*—The Internet Protocol (IP) environment poses two relevant sources of distortion to the speech recognition problem: lossy speech coding and packet loss. In this paper, we propose a new front-end for speech recognition over IP networks. Specifically, we suggest extracting the recognition feature vectors directly from the encoded speech (i.e., the bit stream) instead of decoding it and subsequently extracting the feature vectors. This approach offers two significant benefits. First, the recognition system is only affected by the quantization distortion of the spectral envelope. Thus, we are avoiding the influence of other sources of distortion due the encoding-decoding process. Second, when packet loss occurs, our front-end becomes more effective since it is not constrained to the error handling mechanism of the codec. We have considered the ITU G.723.1 standard codec, which is one of the most preponderant coding algorithms in voice over IP (VoIP) and compared the proposed front-end with the conventional approach in two automatic speech recognition (ASR) tasks, namely, speaker-independent isolated digit recognition and speaker-independent continuous speech recognition. In general, our approach outperforms the conventional procedure, for a variety of simulated packet loss rates. Furthermore, the improvement is higher as network conditions worsen.**

***Index Terms*—Coding distortion, G.723.1, IP networks, IP telephony, packet loss, speech recognition, voice over IP (VoIP).**

I. Introduction

THE RAPID growth of the Internet along with the possibility of endowing Web pages with rich multimedia capabilities are opening a wide variety of e-business opportunities. In other words, the Internet is becoming an ubiquitous vehicle to access a countless number of showcases (the World Wide Web) from every PC, workstation, or cellular phone (the first ones providing Internet access are already available). Such success and popularity are clear proof of the many advantages and potential of the Internet Protocol (IP) as a support for the integration of different kinds of applications and services.

One of the most outstanding examples of the integration support provided by IP is the Internet telephony or voice over IP (VoIP). Contrary to the switched telephone network, IP networks are not intended for transmitting voice. This unsuitability constitutes a very challenging problem; specifically, packet (datagram) loss, delay, and network jitter are the main obstacles for the deployment of VoIP. Nevertheless, despite these drawbacks, IP begins to consolidate as a natural vehicle for the integration of voice and data. There are several reasons for this, but we will highlight the following two:

1) IP is already available in every machine; and
2) this kind of integration can be extended to many other applications (fax, video, shared whiteboards, etc.).

Web-based call centers are one of the most promising VoIP applications [7]. In this context, the ability to provide the client with on-line, cost-effective, friendly spoken interfaces will acutely influence the success of an e-business website. Futhermore, these interfaces will also enable applications ranging from over-the-net dictation to personal assistants, auto attendants, voice dialers, and other computer telephony applications. Nevertheless, the lack of robust speech recognition technologies capable of operating under the adverse conditions imposed by IP networks is currently preventing a wide deployment of spoken language interfaces in Web hosts.

In fact, very limited work has been published on the processing aspects of the integration of automatic speech recognition (ASR) systems on the World Wide Web ([4], [11], and [18] are interesting examples). As far as we know, there has not been any attempt to design specific solutions for the new problems posed by voice transmission over IP networks (from the ASR point of view). In our opinion, the following two obstacles need to be faced:

1) Voice must be encoded for transmission and subsequently decoded at the far end, where it will be recognized. Furthermore, the compression rate is usually high and consequently, the encoding–decoding process causes an impoverishment of the recognition figures.
2) Packet loss severely affects the performance of a speech recognizer, since complete segments of the signal are lost.

In this paper, we begin with a detailed discussion of both problems. Afterwards, in light of this discussion, we propose a new ASR front-end to deal with them. In particular, we suggest performing the recognition from the encoded speech (i.e., from the bit stream) instead of decoding it and subsequently extracting the parameters. We extract and decode only those parameters relevant to the recognition process. In this way, as explained in detail further on, we are preventing part of the coding distortion from influencing the recognizer performance, since the used parameters are directly extracted from the original speech. On the other hand, we cannot avoid the quantization distortion of these parameters, but, as seen ahead, this fact does not particularly affect the recognition performance. We suggested this front-end in the context of digital cellular telephony [6] at the same time as two other similar proposals [8], [2]. Nevertheless, to our knowledge, these ideas are novel in the IP environment.

Manuscript received March 6, 2001. The associate editor coordinating the review of this paper and approving it for publication was Dr. Sadaoki Furui.

The authors are with the Departamento de Tecnologías de las Comunicaciones, Universidad Carlos III de Madrid, Leganés 28911, Spain (e-mail: carmen@tsc.uc3m.es; gallardo@tsc.uc3m.es; fdiaz@tsc.uc3m.es).

Publisher Item Identifier S 1520-9210(01)04318-8.

In order to assess the proposed front-end, we have compared it with the conventional one (i.e., an ASR system operating on the decoded speech signal) using the speech coding standard ITU-G.723.1 [9], under several simulated packet loss rates. In particular, we have tested our procedure in two different ASR tasks, achieving, in general, clear improvements. Furthermore, its benefits increase as the network conditions worsen.

The rest of this paper is organized as follows. Section II presents the specific problems of ASR in the IP environment, discussing the influence of coding distortion and packet loss. Section III reviews the main characteristics of the speech coding algorithm chosen for this work, ITU recommendation G.723.1, as one of the most preponderant codecs for VoIP. Section IV tidily describes our proposal in comparison with the conventional approach. Section V presents the experiments and discusses the results, highlighting the key issues in ASR over IP networks. Finally, some conclusions are drawn and the main lines for future work are outlined in Section VI.

II. Speech Recognition and IP Networks

As stated in Section I, speech recognition technologies are likely going to play an important role in the development of friendly, cost-effective, IP-supported, Web-based services. However, this aim currently poses very challenging technological problems. At this moment, a huge effort is being done in developing solutions at the network and protocol levels. Nevertheless, the network upgrade is very expensive and a long-term solution [7].

In the meantime, present problems should be identified and practical solutions provided. From our point of view, as outlined in Section I, two major difficulties are to be considered: coding distortion and packet loss. The rest of this paper is devoted to these two subjects and their influence on ASR systems.

A. Coding Distortion

Before its transmission over an IP network, the voice signal must be encoded to fit into the available bandwidth. Voice codecs included in the H.323 protocol suite [10] such as G.723.1 and G.729 are the most commonly used ones.

To support some posterior discussions and to gain insight in the actual influence of the coding distortion in ASR tasks, a brief and qualitative description of the main characteristics of the codecs is in order. The G.723.1 and G.729 standard codecs are CELP-type (code excited linear predictive). These codecs achieve low bit rates by assuming a simplified speech production model (known as source-filter model) with negligible interaction between source and filter. The filter is determined on a frame-by-frame basis while the excitation is computed with a higher time resolution (from two to four times per frame, depending of the codec) by means of an analysis-by-synthesis procedure aiming at minimizing a perceptually weighted version of the coding error. As a result, it can be said that these codecs introduce two different types of distortion, namely, that due to the quantization of the parameters to be transmitted and that owing to the inadequacy of the model itself.

Therefore, the waveform, short-time spectrum, and other relevant characteristics of the (encoded and) decoded speech signal are somewhat different from those of the original one.

Very limited work has been reported on the influence of the coding distortion in speech recognition. As far as we know, three papers address this problem directly; the first by Euler and Zinke [5], the second by Dufour *et al.* [3], and the third by Lilly and Paliwal [14]. None of them deals with G.723.1 or G.729, but all of them agree on one general conclusion also applicable to these codecs: even working with matched conditions (i.e., training the system using decoded speech), the speech recognition performances are damaged by codecs working at bit rates under 16 kb/s.

B. Packet Loss

The inadequacy of IP networks for real-time traffic such as voice appears in the form of packet loss, either because the packets are actually lost or because they arrive too late. Depending on the implementation one packet can contain one or more speech frames. For our experiments, one frame per packet is considered.

Obviously, packet loss deteriorates the quality of the decoded speech and several techniques have been proposed to alleviate that problem. According to the taxonomy of error concealment and reparation techniques in [17], one can distinguish between sender-based repair and error concealment techniques by the receiver. The first ones include the traditional *forward error correction* and *interleaving.* Their major inconvenience is the increase in bandwidth requirements. The last ones are independent of the sender and can be further divided into *insertion*, *interpolation*, and *regeneration* techniques. In any case, the objective of these solutions is the recovery of a perceptually acceptable voice waveform; nevertheless, the mismatches between the speech reconstructed in this way and the original one can severely affect the recognition performance. This is the reason why an ASR specific concealment technique such as the one presented in this paper improves the performance of the recognizers.

Again, to support further discussions and to gain insight in the impact of packet loss on recognition performance, we will approximately and briefly describe the general philosophy of the packet concealment techniques implemented in VoIP codecs. When a frame is missing, both the filter (sometimes including a bandwidth expansion) and the excitation of the last correct frame are used instead. The procedure progressively attenuates the excitation until, after a consecutive number of lost frames, the output is finally muted. As a result, single packet loss can be tolerated; however, if packet losses happen in bursts, as usual in the Internet, the consequences can be devastating.

In this paper, we deal with packet losses while bit errors are not considered, since in the VoIP framework bit errors are of only minor importance. Nevertheless, we find it valuable to briefly discuss the essential differences between the impact on ASR performance of packet loss and bit errors. From our experience, we would say that on the one hand, dealing with packet loss is harder because all the information concerning one or more frames is lost; but on the other hand, you can be confident on the received information, and therefore you can rely on it to conceal the missing frames.

III. Speech Codecs for Voice Over IP: G.723.1

Although alternative codecs are emerging, we have investigated the G.723.1 standard codec because, together with the G.729, it is the most widely used in the VoIP environment. Furthermore, G.723.1 seems to be more sensitive to packet loss, mainly due to its relatively slow frame rate (33.3 frames per second). A low frame rate implies that a considerable portion of voice is missing when a packet is lost.

With the purpose of providing a better understanding of the proposed ASR front-end, there follows an outline of the most relevant features of this codec. For a detailed description, we refer the reader to the standard recommendation [9].

The G.723.1 standard is an analysis-by-synthesis linear predictive codec and provides a dual coding rate at 5.3 and 6.3 Kb/s. It is possible to switch between both rates at a frame level and also, an option for variable rate operation is available using voice activity detection (VAD), which compresses the silent portions.

The voice quality offered by G.723.1 can be rated as 3.8 on the M.O.S. scale in 5.3 kb/s mode and 3.9 in 6.3 kb/s mode. Therefore, even though toll quality is claimed, it is obvious that other algorithms provide a slightly better quality: G.729 and G.726 give 4.0 and 4.3, respectively.

G.723.1 uses a frame length L_f of 240 samples (30 ms) and an additional look ahead of 60 samples (7.5 ms), resulting in a total algorithmic delay of 37.5 ms. The frame is divided into four subframes of $L_{sf} = 60$ samples. A window of 180 samples is centered on every subframe and a tenth-order linear prediction (LP) analysis is performed. The prediction coefficients obtained this way are used to implement a short-term perceptual weighting filtering. However, only the coefficients extracted from the last subframe are converted into line spectral pairs (LSP), quantized and sent to the transmission channel. At the decoder, the LSP vector for every subframe of each frame is computed by means of a linear interpolation which involves the current decoded vector and the previous one.

The excitation signal is composed of a periodic and a nonperiodic component. The construction of the periodic component involves the estimation of a pitch lag, and a fifth-order predictor for modeling the long-term correlations among the samples.

The nonperiodic component is computed using different techniques depending on the coding rate used. For the higher rate, 6.3 Kb/s, the encoder uses a multipulse maximum likelihood quantization (MP-MLQ), while at 5.3 Kb/s, it employs an algebraic code excited linear prediction (ACELP) scheme. As will be explained further on, our research focuses on the lower bit rate. In this case, (ACELP) the excitation selection algorithm finds at most four nonzero pulses that can only be allocated at certain fixed positions. Due to these restrictions in the allowed positions, they can be very efficiently encoded.

Another interesting feature of this codec is that it has been designed to be robust against frame erasures. The error concealment strategy, however, must be triggered by an external indication, which can be obtained from the RTP protocol [19]. When the decoder is in concealment mode, it uses the previous LSP vector to produce a prediction of the actual one and generates a synthetic voiced or unvoiced excitation signal based upon a decision taken over the last good frame. The decoded speech is attenuated if bad frames continue to arrive, until it is completely muted after three consecutive losses.

Summing up, a speech frame is encoded through the following parameters:

1) a ten dimension LSP vector, representing its spectral envelope;
2) a pitch lag and a five-dimensional predictor coefficient vector per subframe, representing the periodic fraction of the excitation,
3) a number of positions, the sign, and the gain of the pulses conforming the nonperiodic part of the excitation.

IV. Recognition from Digital Speech

The essential difference between a conventional ASR system and our approach is the source from which the feature vectors are derived. Thus, to assess our proposal, we have tested the two ASR systems that can be observed in Fig. 1. The decoded speech based front-end starts from the decoded speech and proceeds as a conventional ASR system; while the encoded speech based one, does it from a quantized LP spectrum extracted by the G.723.1 encoder. These two different ways of computing the feature vectors are described in more detail in the following subsections.

A. Recognizing Decoded Speech

In this conventional approach, the feature extraction is carried out on the decoded speech signal, which is analyzed once every 10 or 15 milliseconds, employing a 20- or 30-ms analysis Hamming window using the HTK package [22]. Twelve mel-frequency cepstral coefficients (MFCCs) are obtained using a mel-scaled filterbank with 40 channels. Then, the log-energy, the 12 delta-cepstral coefficients and the delta-log energy are appended, making a total vector dimension of 26.

B. Recognizing Digital Speech

Standard speech codecs are completely (bit-level) defined. Therefore, it is possible to selectively access the relevant parameters (from the recognition point of view). The underlying idea here is to feed the speech recognizer with a parameterization directly derived from the digital (encoded) speech representation, i.e., recognizing from digital speech.

This is feasible because, fortunately, as previously noted, the two most preponderant codecs for VoIP, ITU-G.723.1 and ITU-G.729, are CELP-type codecs, and this type of codecs extract and code the appropriate spectral information, from which recognition can be successfully carried out.

One of the aims of our proposal is to reduce the influence of coding distortion on ASR systems performance. Specifically, the spectral envelope derived from the digital speech is the same that would have been obtained from the original speech, except for the quantization. But, as revealed in [20] and confirmed by our experimental results, the quantization distortion does not especially affect the recognition performance. On the other hand, the spectral envelope estimated from the decoded speech could exhibit important differences with respect to the original one,

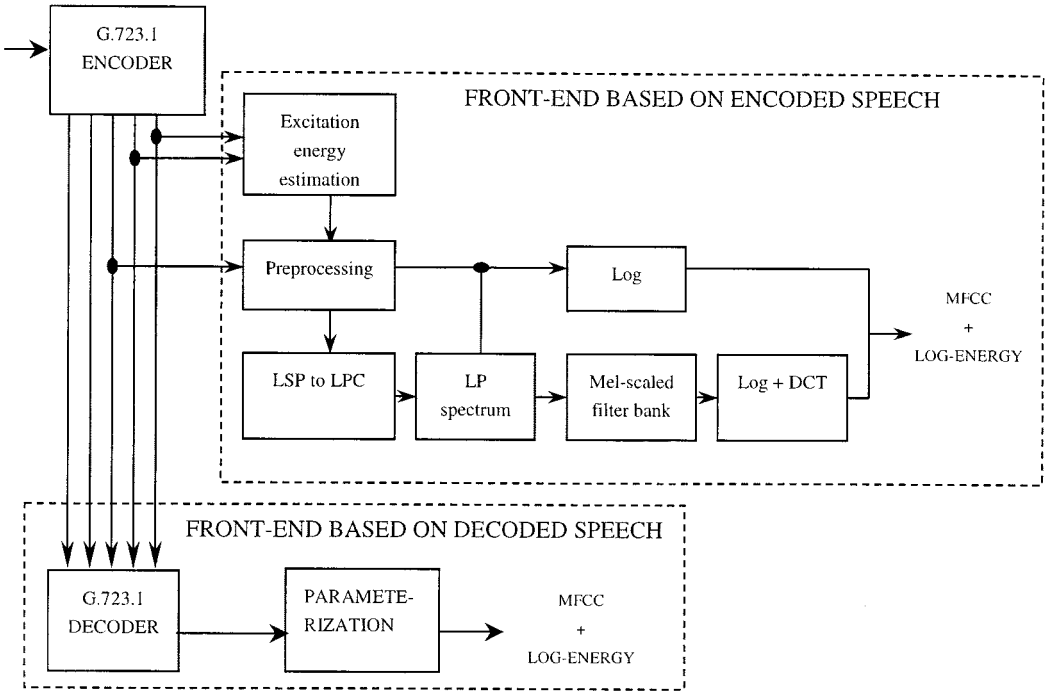

Fig. 1. Parameterization procedures. The lower part of this block diagram illustrates to the steps followed in a conventional approach, i.e., the encoded speech is received at the far end and subsequently decoded before being parameterizated for recognition. The upper part of the diagram represents our proposed procedure, where no decoding is performed. Instead the parameterization is extracted from the quantized LSP coefficients transmitted by the codec subsequently converted into LPC coefficients, LP spectrum, filtered by a mel-scaled filterbank and transformed in MFCC coefficients via discrete cosine transform. By its side, energy is estimated from a subset of the encoded excitation parameters and the aforementioned LP spectrum.

since, as highlighted in Section II-A, the decoded speech is affected by both the quantization distortion of every parameter involved in the speech synthesis and the inadequacies of the source-filter model.

Furthermore, when dealing with packet loss, our front-end reconstructs the missing spectral envelopes from quantized versions of the correctly received ones. This way, it is possible to design procedures that are more effective for the ASR performance. For example, real time requirements on ASR systems are not usually as demanding as on IP telephony. This enables the use of better interpolation procedures, like the one that will be presented in Section V. Moreover, not every parameter needs to be concealed, which prevents bad corrections on unnecessary parameters from adding distortion. On the contrary, when packet loss occurs, the conventional front-end estimates the spectral envelope from the decoded speech, which exhibits degradations due to the effects of the concealment procedures on both spectral envelope and excitation (usually rough, due to the delay time requirements).

Summing up, these are the advantages of the proposed approach.

1) The performance of our system is only affected by the quantization distortion of the spectral envelope and a reduced subset of the excitation parameters. Thus, we are avoiding the distortions due to the quantization of the remaining parameters and possible inadequacies of the source-filter model.
2) When packet loss occurs, our front-end can be more effective since it is not constrained to the error handling mechanism of the codec. In particular, any post-processing will only make use of a trustworthy set of quantized parameters (those extracted from the correctly received frames).
3) The computational effort required is not increased, since the cost of computing the MFCCs from the digital speech is practically equivalent to that of the same task in the conventional front-end; while, in our case, a complete decoding is not necessary.

Nevertheless, it should be admitted that our approach also presents a couple of drawbacks, namely, the front-end should be adapted to the specific codec if we are not willing to accept some mismatch; and, as we will discuss further on, the spectral envelope is available at the frame rate of the codec (which can be too slow). This last is a minor problem which can be easily overcome as will be demonstrated later. In any case, the results shown in the paper indicate that the advantages outweigh the disadvantages.

The block diagram in Fig. 1 illustrates the proposed parameterization procedure compared to the conventional one. Our implementation mixes our own procedures with some facilities of the HTK (HTK Toolkit) package [22]. More precisely, the trans-parameterization (from quantized LSP to MFCC) is described below step by step.

Step 1) For each G.723.1 frame (30 ms of speech), the ten quantized LSP parameters are converted into LP coefficients.

Step 2) A 256-point spectral envelope of the speech frame is computed from the LP coefficients.

Step 3) A filter bank composed of 40 mel-scale symmetrical triangular bands is applied to weight the LP-spectrum magnitude, yielding 40 coefficients, which are

TABLE I
RECOGNITION RATES SHOWING THE INFLUENCE OF SPEECH CODING ON BOTH IDR AND CSR TASKS, FOR THE G.723.1 CODEC. FIRST ROW SHOWS THE REFERENCE EXPERIMENT USING ORIGINAL SPEECH. SECOND AND THIRD ROWS SHOW THE RESULTS WHEN DECODED SPEECH IS INVOLVED, BOTH FOR UNMATCHED (i.e., THE SPEECH MODELS ARE OBTAINED USING ORIGINAL SPEECH BUT THE TEST IS DONE WITH DECODED SPEECH) AND MATCHED CONDITIONS (MODELS ARE TRAINED AND TESTED USING DECODED SPEECH). THE INFLUENCE OF CODING DISTORTION IS NOTICEABLE

Experiment Description (Training-Testing)	IDR Task	95% Confidence Interval	CSR Task	95% Confidence Interval
Original-Original	99.66%	(99.53,99.79)	90.83%	(90.27,91.39)
Original-Decoded	98.98%	(98.76,99.20)	86.28%	(85.61,86.94)
Decoded-Decoded	99.33%	(99.15,99.51)	87.01%	(86.36,87.66)

converted to 12 mel cepstrum coefficients using HTK.

Step 4) The frame energy is estimated as described in the appendix and the log-energy is appended to the feature vector.

Step 5) Dynamic parameters are computed (by HTK) for all the 12 MFCC and the log-energy, making a total vector dimension of 26.

V. EXPERIMENTAL RESULTS

As mentioned in Section III, G.723.1 codec can operate at two different bit rates, namely, 5.3 and 6.3 kb/s. Moreover, the bit rate can be modified at the frame rate. Some preliminary experiments showed that the speech recognition system performance is not very sensitive to the operating bit rate. In other words, from the automatic speech recognition point of view the speech quality at both bit rates is quite similar. Thus, our experiments have focused on the lowest bit rate, 5.3 kb/s, assuming that all of the conclusions can be extended to the highest rate, 6.3 kb/s.

A. Baseline Systems and Databases

In this section, we will present and discuss the experiments carried out in order to compare the proposed front-end with the conventional one in different IP network conditions. For this purpose, we have chosen two different tasks: speaker-independent isolated digit recognition (IDR task) and speaker-independent continuous speech recognition (CSR task).

In order to state the statistical significance of the experimental results shown in the following subsections, we have calculated the confidence intervals (for a confidence of 95%) using the following formula [21, pp. 407–408]:

$$\frac{band}{2} = 1.96\sqrt{\frac{p(100-p)}{n}} \tag{1}$$

where p is the recognition rate for the IDR task or word accuracy for the CSR task and n is the number of examples to be recognized (7920 and 10 288 words for the IDR and CSR tasks, respectively). Thus, any recognition rate in the tables below is presented as belonging to the band $[p-(band/2),\ p+(band/2)]$ with a confidence of 95%.

1) Speaker-Independent Isolated Digit Recognition: For the speaker-independent isolated digit recognition experiments (IDR system), we use a database consisting of 72 speakers and 11 utterances per speaker for the ten Spanish digits. This database was recorded at 8 kHz and in clean conditions. In addition, we have digitally encoded this database using the G.723.1 standard at 5.3 kb/s, so that we have two different databases at our disposal.

Since the databases are quite limited to achieve reliable speaker-independent results, we have used a ninefold cross validation to artificially extend them. Specifically, we have split each database into nine balanced groups; eight of them for training and the remaining one for testing, averaging the results afterwards. In this way, we can include all the 7920 utterances to compute the statistical confidence bands.

The baseline is an isolated-word, speaker-independent HMM-based ASR system developed using the HTK package. Left-to-right HMM with continuous observation densities are used. Each of the whole-digit models contains a different number of states (which depends on the number of allophones in the phonetic transcription of each digit) and three Gaussian mixtures per state.

2) Speaker-Independent Continuous Speech Recognition: The database which we used in our speaker-independent continuous speech recognition experiments is the well-known Resource Management RM1 Database [15], which has a 991 word vocabulary. The speaker-independent training corpus consists of 3990 sentences pronounced by 109 speakers and the test set contains 1200 sentences from 40 different speakers, which corresponds to a compilation of the first four official test sets. Originally, RM1 was recorded at 16 kHz and in clean conditions; however, our experiments were performed using a (downsampled) version at 8 kHz. As in the previous section, we have digitally encoded this database using the G.723.1 standard at 5.3 kb/s.

We have employed context-dependent acoustic models, namely: three-mixture cross-word triphones. The synthesis of unseen triphones in the training set was performed through a decision tree method of state clustering. The standard word-pair grammar was used as the language model.

B. Influence of Coding Distortion on Speech Recognition Performance

We have evaluated the influence of the G.723.1 (5.3 Kb/s) standard speech codec on the performance of the two ASR tasks previously described. Results are shown in Table I, which displays, besides the results achieved in the reference experiment

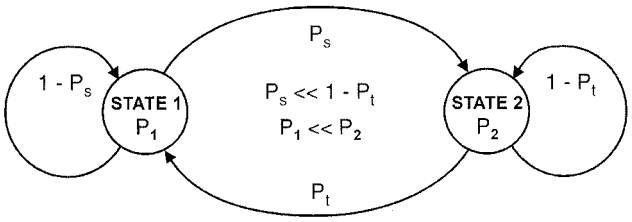

Fig. 2. Model for packet loss simulation. This model is a Markov chain consisting of two states: the first one with a low packet loss rate (P_1) and the second one in which packet loss is highly probable ($P_2 \gg P_1$). Transitions from the state one (good state) to state two (bad state) are modeled through the transition probability P_s. A different probability, P_t, governs the transitions from the bad state to the good one. Bursts are generated by choosing $P_s \ll 1 - P_t$. In these conditions, it is not likely to move from the good to the bad state (P_s), but once the model is in the bad state, is difficult to leave it ($1 - P_t$), thus generating bursts.

using the original speech, two experimental results involving decoded speech, one for matched (training and testing using decoded speech) and another for unmatched conditions (training with original—not encoded—speech and testing with decoded speech).

In both cases (IDR and CSR) the drop in the recognition figures (comparing the reference experiment with any involving encoding speech) are statistically significant, showing that the influence of coding distortion on ASR performance is no longer negligible.

The novel front-end proposed in this paper is aiming at alleviating this influence by circumventing some of the sources of distortion due to the encoding–decoding process.

C. Influence of Packet Loss on Speech Recognition Performance

In order to measure the influence of missing speech packets on the ASR system performance, we have artificially degraded the G.723.1 encoded speech by simulating packet losses produced by the IP channel.

Packet losses encountered in digital transmission over IP are not independent on a frame-by-frame basis, but appear in bursts. Such a channel exhibits memory, i.e., statistical dependence in the occurrence of missing packets. In our case, we have simulated this process using Gilbert's model [12], which represents the behavior of channels with memory in a simple way. Gilbert's model is a Markov chain consisting of two states, as can be observed in Fig. 2: a first one with low packet loss rate (P_1) and a second one in which packet loss is highly probable ($P_2 \gg P_1$). Transitions from the state one (good state) to the state two (bad state) are modeled through the transition probability P_s; thus, $1 - P_s$ represents the probability of remaining in the good state provided we are already there. A different probability, P_t, governs the transitions from the bad state to the good one; with $1 - P_t$ representing the probability of remaining in the bad state. Bursts are generated by choosing $P_s \ll 1 - P_t$. In these conditions, it is not likely to move from the good to the bad state (P_s), but once the model is in the bad state, is difficult to leave it ($1 - P_t$), thus generating bursts.

A recent paper by Borella [1] reports a thorough experimental study about the way in which packet loss occurs in the Internet, focusing on packet lengths and inter-departure times designed for voice traffic according to G.723.1 recommendation. Borella concludes that long-term packet loss rates (PLRs) between 0.5% and 3.5%, with a mean burst length (MBL) of 6.9 packets, can be considered typical. Moreover, approximately 90% of the bursts consist of three packets or less. This fact reveals that some very long bursts occur that significantly contribute to the MBL.

TABLE II
CHARACTERISTICS OF THE IP CHANNELS GENERATED FOR MEASURING THE INFLUENCE OF MISSING SPEECH PACKETS ON THE ASR SYSTEM PERFORMANCE. SPECIFIC VALUES OF THE CHANNEL MODEL PARAMETERS (P_1, P_2, P_s, AND P_t) ARE SHOWN FOR EACH CHANNEL, TOGETHER WITH THE RESULTING PACKET LOSS RATES (PLRS) AND MEAN BURST LENGTHS (MBLS). PLRS AND MBLS ARE NOT THEORETICAL VALUES BUT EXPERIMENTALLY COMPUTED OVER THE DATABASES USED. LAST COLUMN SHOWS THE NUMBER OF LOST FRAMES THAT CONSTITUTE THE 90% OF THE BURSTS; THIS NUMBER HAS BEEN SELECTED ACCORDING TO REAL TRAFFIC RESULTS. CHANNELS A TO E EXPLORE INCREASING PLRS (FROM 0.34% TO 5.83%)

Channel conditions	P_s	P_t	P_1	P_2	PLR	MBL	90% bursts
A	0.001	0.3	0.001	0.85	0.3%	1.76	$\leq$ 3 frames
B	0.002	0.25	0.005	0.85	1.13%	1.61	$\leq$ 3 frames
C	0.005	0.25	0.01	0.85	2.54%	1.62	$\leq$ 3 frames
D	0.005	0.2	0.015	0.85	3.35%	1.63	$\leq$ 3 frames
E	0.01	0.25	0.025	0.9	5.83%	1.70	$\leq$ 4 frames
F	0.01	0.2	0.001	0.9	4.11%	3.23	$\leq$ 7 frames

Although realistic, simulating extremely long bursts is of no use at illustrating the comparisons pursued in this paper, since when a significant part of the speech signal is lost nothing can be done, from the acoustic point of view, to improve the recognition performance. Thus, we have adjusted the Gilbert's model so that 90% of the bursts generated consist of three packets or less, following one of Borella's conclusions. On the other hand, we have cut the likelihood of long bursts by reducing the MBL, although we have also included an example exhibiting a longer mean length.

Likewise, although our experiments focus on PLRs between 0.5% and 3.5%, we have decided to consider also a couple of examples of higher PLRs (up to almost 6%). The main reason to extend the scope of the experiments beyond Borella's typical rates is the high variability exhibited by these experimental measures depending on the number of hops of the particular routing, the geographical locations of the nodes (Paxon [16] claims that Europe suffers considerably higher PLRs than does North America), etc.

Following the above considerations, we have designed six IP channels whose characteristics are listed in Table II. Channels A-E explore increasing PLRs (from 0.34% to 5.83%), always with the 90% of the bursts consisting of three packets or less (except Channel E, which rises this figure to four packets or less). On the contrary, Channel F generates longer bursts (90% of the bursts consist of seven packets or less).

The results, using the G.723.1 codec with its error concealment mechanism, for both the IDR and CSR tasks (for the case in which training and testing is performed with decoded speech) are shown in Table III. We start the discussion with

TABLE III
RECOGNITION RATES SHOWING THE INFLUENCE OF G.723.1 CODING AND PACKET LOSS ON BOTH IDR AND CSR TASKS. THESE EXPERIMENTS WERE CONDUCTED USING DECODED SPEECH FOR BOTH TRAINING AND TESTING

Channel Conditions	IDR Task	95% Confidence interval	CSR Task	95% Confidence interval
-	99.33 %	(99.15,99.51)	87.01 %	(86.36,87.66)
A	99.20 %	(99.00,99.40)	86.64 %	(85.98,87.30)
B	98.91 %	(98.68,99.14)	85.75 %	(85.07,86.43)
C	98.70 %	(98.45,99.95)	84.47 %	(84.18,85.56)
D	98.13 %	(97.83,98.43)	83.66 %	(82.95,84.37)
E	97.53 %	(97.19,97.87)	81.01 %	(80.25,81.77)
F	97.12 %	(96.75,97.49)	81.68 %	(80.93,82.43)

Channels A–E. As expected, the drop in recognition performance increases with the PLR, from 0.13% to 1.81% for the IDR task, and from 0.43% to 6.9% for the CSR task. It is also important to note that the influence is more noticeable in the CSR task.

When the bursts are longer (Channel F) the IDR system seems to be the most impaired. The IDR results for Channel F are poorer than those achieved for Channel E, even though its PLR is lower. However, for the CSR task, Channel F behaves better than Channel E. Very likely, the explanation can be found in the contribution of the language model (only used in CSR), able to conceal some missing information.

These results highlight the remarkable influence of packet loss on the speech recognition accuracy for both tasks, but specially for the CSR one. As will be shown further on, the proposed front-end provides a consistent improvement of the recognition rates in this scenario.

D. Recognition from Digital Speech

Along this subsection we compare the performances achieved by the proposed front-end with those obtained by the conventional one, for the two tasks considered.

It is well known that the recognition figures show a critical dependency on the frame period (the time interval between two consecutive feature vectors). For the IDR task, a 15 ms frame period seems to be appropriate (some experiments were conducted using a frame period of 10 ms, but we did not find any improvement). However, for the CSR task, the frame period should be reduced to 10 ms. In our opinion this is mainly due to the fact that the duration of the acoustic units is shorter in the CSR task (we use word models for IDR and triphones for CSR). As a consequence of this bigger temporal resolution, the acoustic vectors should be extracted at a higher rate. Another (less relevant) argument in the same direction could be that, as some authors state, the speaking rate is usually faster in continuous speech compared to the pronunciations of isolated words.

Our front-end has to deal with the problem of fitting the appropriate frame period, since the G.723.1 standard encodes the LP parameters once every 30 ms. Our first approach to treat this problem entailed replicating the same interpolation scheme used by the G.723.1 decoder. We tried out this solution in the IDR task without the expected success.

TABLE IV
RECOGNITION RATES ACHIEVED FOR THE IDR TASK AND SEVERAL SIMULATED IP CHANNELS. CONFIDENCE BANDS ARE SHOWN IN BRACKETS. AS CAN BE OBSERVED, THE DECREASE IN RECOGNITION RATES DUE TO PACKET LOSSES IS SLOWER FOR THE PROPOSED APPROACH (DIGITAL) THAN IN THE CONVENTIONAL ONE (DECODED). THE CONFIDENCE BANDS ARE OVERLAPPING DUE TO THE SMALL DATABASE USED. A FRAME PERIOD OF 15 ms, WHICH WE HAVE FOUND SUITABLE FOR THIS TASK, IS USED IN BOTH CASES

Channel Conditions	Digital	95% Confidence interval	Decoded	95% Confidence interval
-	99.29 %	(99.04,99.42)	99.33 %	(99.15,99.51)
A	99.17 %	(98.97,99.37)	99.20 %	(99.00,99.40)
B	99.03 %	(98.81,99.25)	98.91 %	(98.68,99.14)
C	98.84 %	(98.60,99.08)	98.70 %	(98.45,98.95)
D	98.59 %	(98.33,98.85)	98.13 %	(97.83,98.43)
E	98.09 %	(97.79,98.39)	97.53 %	(97.19,97.87)
F	97.54 %	(97.20,97.88)	97.12 %	(96.75,97.49)

Given that the standard interpolation did not work out for recognition purposes, we tried to decrease the frame period by means of a smarter interpolation. In fact, the interpolation carried out by the G.723.1 just involves two frames (the current and the previous ones), mainly due to delay constraints; nevertheless, an ASR system can tolerate a delay of a couple of frames (actually, we need such a delay to compute the delta parameters). Following this idea we have tested out a band-limited interpolation FIR filter on the LSP coefficients to obtain a frame period of 15 ms (for the IDR task) or 10 ms (for the CSR task). The interpolation filter uses the nearest four (two of each side) nonzero samples.

Tables IV and V show the results achieved by our front-end (labeled as *Digital*) in comparison with the conventional one (labeled as *Decoded*) for the IDR and CSR tasks, respectively. In order to address the statistical significance of the experimental results, the confidence intervals calculated for a confidence of 95% are displayed along with the recognition rates. Finally, the performances of both systems have been evaluated for the six IP channels described in the previous subsection. In any case, the training is performed on *clean* (not affected by packet loss) speech.

For the IDR task, it seems clear that the proposed front-end performs slightly better than the conventional approach. The results are just equivalent to those achieved by the conventional method when no packet loss is considered or the PLR is very low (Channel A). Nevertheless, for Channels B-E, i.e., for PLRs between 1.13% and 5.83%, our front-end provides better results, with the improvement increasing with the PLR [from 0.12% (Channel B) to 0.57% (Channel E)]. The results obtained for Channel F are also favorable to our approach. Although the PLR in this last case (4.11%) is higher than that of the Channel D

TABLE V
RECOGNITION RATES ACHIEVED FOR THE CSR TASK AND DIFFERENT SIMULATED IP CHANNELS. THE DECREASE IN RECOGNITION RATES DUE TO PACKET LOSSES IS SLOWER FOR THE PROPOSED APPROACH (DIGITAL) THAN IN THE CONVENTIONAL ONE (DECODED). EVEN FOR LOW PACKET LOSS RATE CHANNELS (e.g., A AND B), THE DIGITAL APPROACH IS STILL ADVANTAGEOUS. A FRAME PERIOD OF 10 ms, WHICH WE HAVE FOUND SUITABLE FOR THIS TASK, IS USED IN BOTH CASES

Channel Conditions	Digital	95% Confidence interval	Decoded	95% Confidence interval
-	88.33 %	(87.71,88.95)	87.01 %	(86.36,87.66)
A	88.25 %	(87.63,88.87)	86.64 %	(85.98,87.30)
B	87.46 %	(86.82,88.10)	85.75 %	(85.07,86.43)
C	86.81 %	(86.17,87.46)	84.47 %	(84.18,85.56)
D	86.09 %	(85.42,86.76)	83.66 %	(82.95,84.37)
E	83.98 %	(83.27,84.69)	81.01 %	(80.25,81.77)
F	83.96 %	(83.25,84.67)	81.68 %	(80.93,82.43)

(3.35%) the performance improvement is smaller, since the likelihood of long bursts, devastating for both approaches, is higher, thus leaving less room for improvement.

In any case, it should be noted that the database used for the IDR task is not large enough to guarantee the statistical relevance of the improvements with a confidence of 95%.

For the CSR task, the proposed front-end always provides better results than the conventional one. Moreover, the improvements are statistically significant for all of the IP channels considered. The improvements in this case exhibit the same trend that for the IDR task, i.e., they increase with the PLR. However, in this case, the improvements are considerably higher, starting from a 1.52% (for no packet losses), and reaching a 3.67% (for the Channel E). Finally, the same comment about the Channel F for the IDR system results applies to the CSR system.

VI. CONCLUSIONS AND FURTHER WORK

After reviewing the new difficulties faced by speech recognition technologies in the VoIP environment, namely, speech coding distortion and packet loss, we have proposed a new front-end for speech recognition on IP networks. In particular, we suggest performing the recognition from the encoded speech (i.e., the bit stream) instead of decoding it and subsequently proceed to the recognition. In this way, we are circumventing the influence on the recognizer of some sources of distortion due to the encoding–decoding process. Furthermore, when packet loss occurs, our front-end becomes more effective, since it is not constrained to the error-handling mechanism of the codec.

We have evaluated our front-end and compared it to the conventional approach in two ASR tasks, namely, speaker-independent IDR, and speaker-independent CSR. The comparison has been conducted in several simulated packet loss conditions derived from real voice traffic measurements over the Internet.

We have identified the frame rate of the speech codec as a key issue to be considered. In particular, the G.723.1 codec encodes and transmits the spectral envelope every 30 ms and it is necessary to increase the rate at which the spectral information is fed to the speech recognizer to achieve the best performance. For this purpose, we have proposed an interpolation scheme which has proved to be very effective.

From our results, the following conclusions can be drawn. First, for the IDR task, our approach is superior only for the channels with PLRs of 1.13% and higher (although the database is not large enough to guarantee the statistical significance with a confidence level of 95%). Second, for the CSR task, the proposed front-end provides significant improvements for all of the IP channels considered, even when low PLRs or no packet loss are considered.

Third, for both tasks, the decrease in the recognition rates due to packet losses is slower in our approach than in the decoded one. In other words, the worse the conditions of the IP network are, the higher the benefits of our technique become. Therefore, it can be concluded that the proposed approach is much more robust than the conventional one. In our opinion, this is due to that any kind of processing intending to conceal the missing information will be supported in a trustworthy set of quantized parameters (those extracted from the correctly received frames). Furthermore, it is not constrained to the error handling mechanism of the codec.

This paper has focused on ITU-G.723.1 speech codec; however, this approach could be also easily extended either to other standards codecs (like G.729), or even to proprietary ones, since, in every case, low bit rate codecs typically used in VoIP systems are CELP-type and, consequently, encode and transmit the spectral envelope of the speech signal. Furthermore, it is expected that the proposed method would attain even better results working with the standard G.729, since this codec uses a 10-ms frame rate, thus avoiding the need of interpolation. We leave these experiments for further work.

Finally, we feel that there is room to investigate more elaborated ways of reconstructing the missing information due to packet loss.

APPENDIX
FRAME ENERGY ESTIMATION

Almost every speech recognizer includes the energy of the speech signal in the parameter vector. However, the G.723.1 standard does not encode the energy as a separate parameter and therefore, it should be computed from some of the encoded parameters.

We have calculated the mean power of every subframe as follows. Modeling the excitation $e[n]$ in every subframe as a zero-mean white Gaussian noise, the mean power of the synthesized speech in the corresponding subframe can be computed as follows:

$$\sigma_x^2 = \sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(\Omega)|^2 \, d\Omega \tag{2}$$

where σ_e^2 is the variance of the excitation and $H(\Omega)$ is the frequency response of the synthesis filter.

Let $\hat{\sigma}_x^2[k, i]$ denote the estimated mean power of the subframe i ($0 \leq i \leq 3$) of the frame k. Following (1), $\hat{\sigma}_x^2[k, i]$ can be calculated as

$$\hat{\sigma}_x^2[k, i] = \hat{\sigma}_e^2[k, i] \cdot \hat{E}_h[k, i] \tag{3}$$

where $\hat{\sigma}_e^2[k, i]$ and $\hat{E}_h[k, i]$ represent the estimations of the excitation variance and the contribution of the synthesis filter, respectively. In the following exposition, the frame and subframe indexes, k and i, will be dropped for simplicity and recalled appropriately when necessary.

Starting with the filter contribution, $\hat{E}_h$ can be easily obtained approximating the integral of (1) by the following sum involving the 256-point spectral envelope calculated from the LSP coefficients (Step 2 of the trans-parametrization described in Section IV-B):

$$\hat{E}_h = \frac{1}{N}\sum_{r=0}^{N-1}\left|H\left(\frac{2\pi}{N}r\right)\right|^2 \tag{4}$$

where $N = 256$ in our case.

Before exposing how $\hat{\sigma}_e^2$ has been estimated, a description of the excitation encoding procedure performed by the G.723.1 is necessary: the excitation signal $e[n]$ is computed as the sum of two vectors: the adaptive codebook excitation vector $u[n]$ and a contribution from a fixed codebook $v[n]$.The adaptive codebook contribution comes from a fifth-order pitch prediction defined as follows:

$$u[n] = \sum_{j=0}^{4}\beta_{ij}e'[n+j] \qquad 0 \le n < L_{sf} \tag{5}$$

where L_{sf} is the subframe length; $\beta_{ij}$is the jth coefficient of the pitch predictor for the ith subframe ($0 \le i \le 3$); and $e'[n]$ is a signal constructed as follows:

$$\begin{aligned} e'[0] &= e[-L_i - 2] \\ e'[1] &= e[-L_i - 1] \\ e'[n] &= e[(n \bmod L_i) - L_i - 2], \qquad 2 \le n \le L_{sf} + 3 \end{aligned} \tag{6}$$

with L_i being the pitch lag obtained for the same subframe.

For the estimation of the variance of the excitation, we will assume that the adaptive and fixed codebook contributions are uncorrelated and thus

$$\hat{\sigma}_e^2 = \hat{\sigma}_u^2 + \hat{\sigma}_v^2 \tag{7}$$

where $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ are the estimations of the adaptive and fixed codebook contributions, respectively.

Recalling the generation procedure of the adaptive codebook [from (4)], an estimate of the adaptive contribution can be easily obtained as

$$\hat{\sigma}_u^2[k, i] = \frac{1}{L_{sf}}\sum_{n=0}^{L_{sf}-1}\left(\sum_{j=0}^{4}\beta_{ij}e'[n+j]\right)^2 \approx \sum_{j=0}^{4}\beta_{ij}^2\hat{\sigma}_{e'}^2[j] \tag{8}$$

where $\hat{\sigma}_{e'}^2$ is the variance of $e'[n]$ defined in (5), and the cross products of the quadratic sum have been neglected.

Now, we can obtain $\hat{\sigma}_v^2$ from the number of pulses N_p that conform the nonperiodical excitation ($N_p = 4$) and the gain G applied to the fixed codebook

$$\hat{\sigma}_v^2 = \frac{N_p}{L_{sf}} \cdot G^2. \tag{9}$$

Finally, for the optimal performance of our system, the pre-processing depicted in Fig. 1 must be identically applied to both the LPSs and the excitation energy. Consequently, we only need to estimate the variance of the excitation once per frame. In particular, we have used $\hat{\sigma}_x^2[k, 3]$ as the energy estimation since the subframe 3 is aligned with the decoded LSP vector.

This alignment is evident if we look at the LSP interpolation formula

$$\hat{\mathbf{p}}[k, i] = \begin{cases} 0.75\hat{\mathbf{p}}[k-1, 3] + 0.25\hat{\mathbf{p}}[k, 3], & i = 0 \\ 0.50\hat{\mathbf{p}}[k-1, 3] + 0.50\hat{\mathbf{p}}[k, 3], & i = 1 \\ 0.25\hat{\mathbf{p}}[k-1, 3] + 0.75\hat{\mathbf{p}}[k, 3], & i = 2 \\ \hat{\mathbf{p}}[k, 3], & i = 3 \end{cases} \tag{10}$$

where $\hat{\mathbf{p}}[k, i]$ is the decoded LSP vector for the ith subframe of the kth.

Note, however, that the excitation is never re-synthesized and that, for the energy estimation, we simply decode the gain G, the pitch lag L_i, and the pitch predictor coefficients β_{ij}.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] M. S. Borella, "Measurement and interpretation of internet packet loss," *J. Commun. Network.*, vol. 2, no. 2, pp. 93–102, June 2000.
[2] S. H. Choi, H. K. Kim, and H. S. Lee, "Speech recognition using quantized LSP parameters and their transformations in digital communication," *Speech Commun.*, vol. 30, pp. 223–233, Apr. 2000.
[3] S. Dufour, C. Glorion, and P. Lockwood, "Evaluation of the root-normalized front-end (RN-LFCC) for speech recognition in wireless GSM network environments," in *Proc. ICASSP*, vol. 2, Atlanta, GA, 1996, pp. 77–80.
[4] V. V. Digalakis, L. G. Neumeyer, and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the world wide web," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 82–90, Jan. 1999.
[5] S. Euler and J. Zinke, "The Influence of speech coding algorithms on automatic speech recognition," in *Proc. ICASSP*, vol. 1, Australia, 1994, pp. 621–624.
[6] A. Gallardo-Antolín, F. Díaz-de-María, and F. Valverde-Albacete, "Recognition from GSM digital speech," in *Proc. ICSLP*, vol. 4, Sidney, Australia, 1998, pp. 1443–1446.
[7] M. Hassan, A. Nayandoro, and M. Atiquzzaman, "Internet telephony: Services, technical challenges and products," *IEEE Commun. Mag.*, vol. 38, pp. 96–103, Apr. 2000.
[8] J. M. Huerta and R. M. Stern, "Speech compression from GSM codec parameters," in *Proc. ICSLP*, vol. 4, Sidney, Australia, 1998, pp. 1463–1466.
[9] "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s,", ITU-T Recommendation G.723.1, Mar. 1996.
[10] "Packet-based multimedia communication systems,", ITU-T Recommendation H.323, Feb. 1998.
[11] , L. Julia, A. Cheyer, L. Neumeyer, J. Dowding, and M. Charafeddine. [Online]. Available: http://www.speech.sri.com/demos/atis.html
[12] L. N. Kanal and A. R. K. Sastry, "Models for channels with memory and their applications to error control," *Proc. IEEE*, vol. 66, pp. 724–744, Jul. 1978.
[13] T. J. Kostas and M. S. Borella *et al.*, "Real-time voice over packet-switched networks," *IEEE Network*, vol. 12, pp. 18–27, Jan./Feb. 1998.
[14] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proc. ICSLP*, vol. 4, Philadelphia, PA, 1996, pp. 2344–2347.
[15] *The Resource Management Corpus (RM1)*, Distributed by NIST, 1992.
[16] V. Paxon, "Measurements and analysis of end-to-end internet dynamics," Ph.D. dissertation, Univ. of California, Berkeley, 1997.

[17] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet-loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, pp. 40–48, Aug. 1998.
[18] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments," in *Proc. ICASSP*, Seattle, WA, 1998, pp. 977–980.
[19] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP A transport protocol to real-time applications," *Internet Engineering Task Force*, July 2000.
[20] R. Tucker, T. Robinson, J. Christie, and C. Seymour, "Compression of acoustic features—Are perceptual quality and recognition performance uncompatible goals?," in *Proc. Eurospeech*, vol. 5, Budapest, Hungary, 1999, pp. 2155–2158.
[21] N. A. Weiss and M. J. Hasset, *Introductory Statistics*, 3rd ed. Reading, MA: Addison-Wesley, 1993.
[22] S. Young *et al.*, *HTK—Hidden Markov Model Toolkit (ver 2.1)*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

Carmen Peláez Moreno received the telecommunications engineering degree from the University of Navarre, Spain, in 1997. She completed her final year project at the University of Westminster, London, U.K., and is currently pursuing the Ph.D. degree in the Department of Communication Technologies, University Carlos III, Madrid, Spain. Her research interests include speech coding and recognition, wireless networks, VoIP, QoS, and neural networks.

Ascensión Gallardo-Antolín received the telecommunication engineering degree in 1993 from Universidad Politécnica de Madrid, Spain, where she began pursuing the Ph.D. degree in robust speech recognition.

Since 1997, she has been Assistant Researcher and Professor at University Carlos III, Madrid. Her research interests are in speech recognition, spoken information retrieval, and dialog and signal processing for multi-media human–machine interaction. She has coauthored several communications to international conferences, mainly in speech recognition. She has participated in several research projects including some for the Spanish Council on Science and Technology and the UE.

Fernando Díaz-de-María (S'92–M'96) received the telecommunication engineering degree and the Ph.D. degree from the Universidad Politécnica de Madrid, Spain, in 1991 and 1996, respectively.

He is an Associate Professor at the Department of Communication Technologies, University Carlos III, Madrid, Spain. He is currently Director of the Ph.D. program for Telecommunication Technologies and Associate Director for Telecommunication Engineering at the same institution. His primary research centers on speech coding and speech recognition. Other related research is on nonlinear signal processing (mainly using artificial neural networks) and multimedia.