SVM-BASED LOST PACKETS CONCEALMENT FOR ASR APPLICATIONS OVER IP

C. Peláez-Moreno, A. Gallardo-Antolín, E. Parrado-Hernández and F. Díaz-de-María

Dpto. de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, SPAIN

ABSTRACT

Voice over IP is becoming very popular due to the huge range of services that can benefit from integrating different media (voice, audio, data, etc.). Besides, voice-enabled interfaces for those services based on speech recognition are being very actively researched. Nevertheless, as it has been already pointed out ([8]), the impact of packet losses on speech recognizers is very significant. In this paper, we have compared the usual concealment method (typically implemented by speech codecs) with an SVM-based one. The proposed method is substantially better in terms of mean-square error when isolated packet losses are considered. On the other hand, preliminary ASR experiments in a more realistic environment (bursty packet losses) show just a slight enhancement of the recognition accuracy. We expect significant improvements by adapting the SVM-based method to the bursty losses.

1. INTRODUCTION

During the last years, the interest on Voice over IP (VoIP) networks has considerably grown. Besides the benefits encountered in the IP telephony itself, the transmission of voice over IP is called to play a relevant role as a vehicle for providing voice-enabled interfaces for WWW-based services.

Nevertheless, several problems need to be addressed before speech technologies reach a wide deployment in this context since, nowadays, the quality of voice over IP is notably worse than that obtained through the traditional and voice-oriented telephone network.

The main causes of this lower quality are well known: the speech coding distortion, the unpredictable delay of the packets that carry the coded voice frames (jitter) and the packet losses that occur due to congestions in the network nodes.

In this paper we focus on packet losses and its influence on Automatic Speech Recognition (ASR) systems; one of the speech technologies involved in the creation of voice-enabled WWW interfaces. In this context and simulating packet losses, we assess the performance of usual packet concealment methods (those built in speech coders) and propose a new one, based on a Support Vector Machine (SVM).

When packet losses occur, current speech coders just repeat the last available parameters (at least, those related to the spectral envelope of the speech signal which are the ones containing the main information for ASR). This simple solution, though satisfactory for human beings recognizers, causes an important impoverishment of automatic recognition performance (e.g. [4] and [8]) and consequently a lack of reliability on the corresponding voice-enabled interfaces.

In particular, we propose to estimate the corresponding spectral parameters from previously (or even subsequently) received ones using an SVM-based regressor.

The paper is organized as follows: in section 2 we introduce the problem of packet losses in the IP network. Next, section 3 is devoted to the explanation of the SVM-based procedure we propose for parameter concealment, followed by the description of the experiments we have conducted for assessing our proposal. Finally, we draw some conclusions and outline some further work.

2. ASR OVER IP NETWORKS: DEALING WITH PACKET LOSSES

When a packet is lost, the information concerning the corresponding speech frame will not be available in the ASR system. For mitigating the impact of this loss on recognition performance, we propose making an estimation of the spectral parameters from previously (or even subsequently) received ones.

Speech coders implement a very simple estimation procedure consisting on repeating the last received spectral parameters. This is a good solution for speech coding for several reasons: 1) it does not incur in any extra computational cost; 2) it does not cause additional coding delay; and 3) since the actual bandwidth of spectral parameters is remarkably low (around 7 Hz [6]), just a very slight variation can be expected between two consecutives frames.

Nevertheless, as evidenced by previous ASR results considering packet losses ([8]), this concealment procedure is not suitable for recognition purposes. Furthermore, computational and delay constraints are not

so tight in recognition applications: on the one hand, recognition is itself a very time consuming task and on the other, the computation of commonly employed dynamic parameters (delta) produces at least one or two frames delay.

In this paper we propose a more versatile estimation procedure which takes advantage of the relaxation of the computational and delay constraints –posed by speech coders– not concerning ASR applications.

For these reasons, the proposed technique can not be used in a typical scenario such as that illustrated in Figure 1. In such a case, when a packet is lost, the standard concealment mechanism commonly implemented in speech coders inevitably starts.

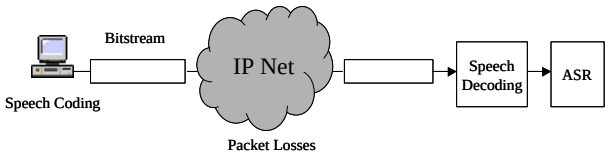

Figure 1. – ASR over IP networks

Therefore, to apply these ideas, we should turn towards somewhat different approaches. There are two possibilities: bitstream-based ASR ([4] or [8], for example) or distributed ASR [2] illustrated in Figure 2 and Figure 3, respectively.

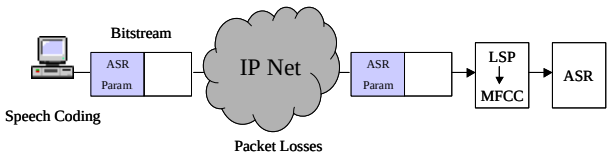

Figure 2.- Bitstream-based remote ASR

In a bitstream-based ASR, the recognition is performed from the encoded speech (i.e., from the bit stream) instead of decoding it and subsequently extracting the recognition parameters. Only those parameters relevant to the recognition process are extracted and decoded. This approach has been shown to be more robust that the conventional approach depicted in Figure 1 both for mobile of IP networks ([4], [8]). Proceeding this way, we have access to the sequence of spectral parameters (typically LSP) which enables us to apply an ASR-specific concealment technique.

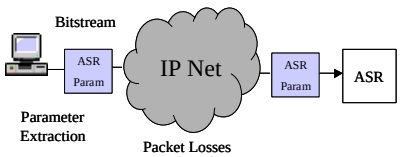

Figure 3.- Distributed ASR

Distributed ASR is illustrated in Figure 3. In this case, the speech signal is not encoded and transmitted, but just the parameters useful for ASR. Thus, part of the recognition process is performed at the terminal (namely, the parameterization), and the remaining at the ASR server. The advantages of this approach rely on the fact that the bandwidth required to transmit the recognition parameters is very small, while the computational effort needed for the parameter extraction is not so high.

Though in this paper we have concentrated our efforts on the bit-stream based approach, it is noteworthy that the distributed one also allows us to apply an ASR-specific concealment technique like the one we propose since we have direct access to spectral parameters (MFCC in this case).

3. A SVM-BASED LOST PACKET CONCEALMENT TECHNIQUE

3.1 Support Vector Machines (SVM)

We use Gaussian Kernel Support Vector Machines (SVMs) regressors [10] to reconstruct the useful (from the recognition point of view) information of the missing frames. These machines seek to determine a function $f(\mathbf{x})$ that, for each point $\mathbf{x_i}$ of the training set, verifies $|f(\mathbf{x_i}) - y_i| < \varepsilon$. So to speak, the algorithm fits a hosepipe of radius ε to the data. The smoothness of the estimated function is controlled by means of allowing some of the data points to remain outside the pipe.

The regressor is the result of an optimization problem applied to an RBF Neural Network architecture whose nodes are some critical input data named Support Vectors (SVs):

$$f(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b \qquad (1)$$

where $\mathbf{x_i}$, with $1 \le i \le M$, are the SVs, M is the number of SVs, k is a Gaussian kernel and α_i and b are the coefficients of the linear combination, that result from the optimization problem. The SVM automatically determines the SVs, so that there is no need to fix *a priori* the architecture of the RBF network.

The experiments carried out in this paper have been run with the MySVM implementation of SVM, available in [9]. The parameters of the Gaussian kernel have been determined experimentally.

3.3 An SVM-based concealment technique

As we have already mentioned, we will focus on the bit-stream based ASR, though exactly the same procedures can be applied to distributed ASR.

Once the bit-stream reaches the ASR server, some of the parameters are extracted to feed the automatic speech recognizer. In particular, we extract 10 LSPs –Line Spectral Pairs– which represent the spectral envelope, and the energy of the corresponding frame. Later on, the LSPs coefficients will be transformed into cepstral parameters to proceed with the HMM-based ASR [8].

Initially, we had considered the evolution of each LSP as a time series. An SVM was trained for each parameter to predict the following value from previous ones. We achieved significant estimation improvements with respect to repetition. Nevertheless, recognition performance did not improve at all. In our opinion, the reason is clear: the spectral information is embedded, not only in the absolute LSP values, but in the relative differences between them. Therefore, considering individually each LSP coefficient trajectory does not seem a suitable approach.

To improve our results we have estimated each LSP from the whole previous and posterior LSP vectors (as illustrated in Figure 4). In this way we are implicitly considering their relative values and, furthermore, we exploit the existing correlation among them.

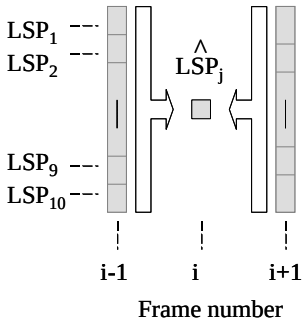

Figure 4.- Vectorial and bidirectional estimation scheme

It is important to notice that the computational cost of SVM-based reconstruction is not very significant in the ASR context.

5. EXPERIMENTS AND RESULTS

5.1. Experimental Set-up

Parameters extraction. The database which we have used in our estimation and speaker-independent continuous speech recognition experiments is the well-known Resource Management RM1 Database [7], which has a 991 words vocabulary. The speaker-independent training corpus consists of 3,990 sentences pronounced by 109 speakers and the test set contains 1,200 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 (down-sampled) version at 8 kHz.

The feature extraction is carried out analyzing the speech signal once every 10 ms employing a 20 ms analysis Hamming window using the HTK package [11]. Ten Linear Prediction (LP) coefficients and an energy parameter are subsequently computed for each of these analysis windows. Finally, the LSP coefficients are obtained from the LP coefficients using the usual transformation (see, for example [5]).

SVM training. With the purpose of training the SVM regressor (SVR) and testing its performance, we have chosen two subsets from the RM1 database training corpus for the training and validation of the SVM, respectively. Thus, the SVM training set consists of 109 sentences, each of which belongs to one of the 109 speakers, yielding a total of 32,232 examples (one for each speech frame). Similarly, the SVM validation set provides 36,043 examples.

Packet loss simulation. In order to measure the influence of missing speech packets on the ASR system performance, we have simulated packet losses produced by the IP channel.

Packet loss 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 [3], which represents the behavior of channels with memory in a simple way.

Channel	P_S	P_T	P_1	P_2	PLR	MBL
A	0.001	0.1	0.001	0.85	0.76%	2.81
B	0.002	0.1	0.005	0.85	1.27%	1.81
C	0.003	0.3	0.01	0.75	2.84%	2.09
D	0.005	0.1	0.01	0.85	4.14%	2.39
E	0.005	0.1	0.015	0.85	10.41%	2.80

Table 1. Simulated IP channel conditions.

According to the experiments performed by Borella [1] about the transmission of VoIP over the public Internet we have designed five different channels exhibiting a range of Packet Loss Rates (PLRs) and Mean Burst Lengths (MBL) as can be observed in *Table 1*. (See [8] for more details). We have taken into account the fact that the 10 ms frame we are considering would be rather packetized in pairs or trios.

5.2. Estimation experiments

To gain some insight on the problem, we have compared the mean-square error (MSE) for both the repetition procedure and the SVM regressor and the

results, highly favourable to the later, are shown in *Table 2*.

LSP	Substitution by repetition	SVM regressor
1	0.2787	0.1352
2	0.2172	0.0974
3	0.1890	0.0986
4	0.1260	0.0672
5	0.1400	0.0674
6	0.1530	0.0773
7	0.1458	0.0722
8	0.2002	0.0994
9	0.2291	0.1197
10	0.2620	0.1462

Table 2. Mean-square error in the repetition and SVM approaches.

It is worth noting that this test validates the performance of the SVM regressor when isolated errors occur.

5.3. Recognition experiments

Finally, we have tried our proposal using the channels we have described in section 5.1. As can be seen in *Table 3*, slight improvements are obtained using our proposed method except for the E channel which presents an extremely high PLR.

Channel	Substitution by repetition	SVM regressor
A	89.93	89.94
B	89.86	89.87
C	89.26	89.37
D	89.01	89.05
E	86.05	86.00

Table 3. Word error rate (%) of the HMM-based recognizer.

However, there is an imbalance between what can be expected from the errors displayed in *Table 2* and the recognizer performance. This is due, in our opinion, to the substantial differences between the first (MSE computation) and the second (ASR) experiments, since for the last one bursty packet losses are considered. The SVMs has been trained to deal with isolated errors and are faced to bursty ones. In any case, these preliminary results indicate that appropriate modifications can lead to important improvements.

6. CONCLUSIONS AND FURTHER WORK

In this paper, we have considered the problem that packet losses pose over ASR performance when voice travels over IP networks. We have compared the usual concealment method employed by speech coders with a more versatile method based on SVM regressors.

Our experiments show that while obtaining remarkably good results in terms MSE, the ASR performances just slightly improve. Our method is suboptimal for the bursty packet losses taking place in the public internet and should be improved to properly deal with bursts. Furthermore, we believe this technique can be improved in more ways, for example: SVM parameters tuning or individual procedures for the energy parameter.

Another important application of this method, namely distributed ASR, is some of the further work we are considering.

7. ACKNOWLEDGMENTS

This work has been partially supported by Spain CICYT grant TIC-1999-0216 and Spain CAM-07T-0018-2000

8. REFERENCES

[1] Borella, M. S.: Measurement and Interpretation of Internet Packet Loss, Journal of Communications and Networking, vol. 2, no. 2, pp. 93-102, (2000)

[2] Digalakis, V.V., Neumeyer, L.G. and Perakakis, M., "Quantization of Cepstral Parameters for Speech Recognition Over the World Wide Web," IEEE Journal on Selected Areas in Communications, vol. 17, no. 1, pp. 82-90, Jan. 1999.

[3] Kanal L. N. and Sastry, A. R. K., "Models for Channels with Memory and Their Applications to Error Control," *Proceedings of the IEEE*, Vol. 66, pp. 724-744, Jul. 1978.

[4] Kim, H. K., Cox, V.: A bitstream-based front-end for wireless speech recognition on IS-136 communications system, IEEE Transactions on Speech and Audio Processing, vol. 9, no. 5 (2001)

[5] Kondoz, A. M.: Digital speech: coding for low bit rate communication systems, Ed. John Wiley & Sons, (1996)

[6] Nadeu, C, Pachès-Leal, P. and Juang, B.-H.: Filtering the time sequences of spectral parameters for speech recognition, Speech Communication 22, pp. 315-322 (1997)

[7] National Institute of Standards and Technology (NIST) (distributor): The Resource Management corpus part 1 (RM1) (1992)

[8] Peláez-Moreno, C., Gallardo-Antolín, A., Díaz-de-María, F.: Recognizing Voice over IP networks: a Robust Front-End for Speech Recognition on the WWW, IEEE Trans. on Multimedia, vol. 3, no. 2, pp. 209-18 (2001)

[9] Rüping, S.: mySVM-Manual. University of Dortmund, Lehrstuhl Informatik 8, http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/ (2000)

[10] Schölkopf, B. and Smola, A.J.: Learning with Kernels. MIT Press, Cambridge MA, (2002)

[11] Young, S. et al: HTK-Hidden Markov Model Toolkit (ver. 3.0), Cambridge University, 2000.