An Application of SVM to Lost Packets Reconstruction in Voice-Enabled Services

C. Peláez.Moreno[1], E. Parrado-Hernández[1], A. Gallardo-Antolín[1], A. Zambrano-Miranda[1] and F. Díaz-de-María[1]

Dpto. Teoría de la Señal y Comunicaciones, Escuela Politécnica Superior-Univ. Carlos III de Madrid, Avda. Universidad, 30, 28911 Leganés, Madrid, Spain.
carmen@tsc.uc3m.es,
http://www.tsc.uc3m.es

Abstract. Voice over IP (VoIP) is becoming very popular due to the huge range of services that can be implemented by integrating different media (voice, audio, data, etc.). Besides, voice-enabled interfaces for those services are being very actively researched. Nevertheless the impoverishment of voice quality due to packet losses severely affects the speech recognizers supporting those interfaces ([8]). In this paper, we have compared the usual lost packets reconstruction method with an SVM-based one that outperforms previous results.

1 Introduction

With the consolidation of IP networks as the preferred and most commonly implemented packet-based networks, the interest on the transmission of voice over these networks has considerably grown. One of its attractiveness is the possibility of using it as a vehicle for providing voice-enabled interfaces for WWW services.

However, to attain this goal, several problems need to be addressed, since the quality of voice over IP is considerably worse than the obtained using the traditional and voice-oriented telephone network. The main causes of this impoverishment of the quality are well known: the need of a severe compression of the speech signal, the unpredictable delay of the packets that carry the individual voice frames 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 both speech quality and subsequent Automatic Speech Recognition (ASR) in voiced-enable services. In this context, assuming that each packet bears a coded speech frame and simulating packet losses according to actual measurements of the Internet behavior ([2]), we asses the performance of usual reconstruction methods and propose a new one, based on a Support Vector Regressor (SVR).

When packet losses occur, current speech codecs just repeat the last available parameters. This simple solution, as stated by previous works, leads to an impoverishment of recognition performance (e.g. [8], [3]) and consequently to less reliable voice-enabled interfaces.

The paper is organized as follows: in section 2 we present the problem of packet losses in IP networks. Next, section 3 is devoted to the explanation of the SVM-based procedure we propose for parameter reconstruction. In section 4, we describe the experiments we have conducted for assessing our proposal. Finally, we draw some conclusions and outline some further work.

2 ASR over the WWW: facing the packet losses

To gain insight on the actual causes of this loss of recognition performance we have conducted the following experiment. On the one hand, we have obtained the recognition parameters from the original speech. On the other hand, we have simulated packet losses –according to real measurements due to Borella [2]- and generated another set of parameters in which the lost parameters are substituted by the previous ones. Finally, we have computed the mean square reconstruction error.

Surprisingly, as will be shown in section 4, the average errors are very low. However, the recognition performance significantly decreases. In fact, repeating the last parameters is actually a good solution since, as evidenced by Nadeu et al. [5], the bandwidth of the spectral parameters is extremely low, i.e., its time evolution is very slow.

The previous considerations allow us to conclude that slight improvements on reconstruction errors are likely going to lead to significant improvements on ASR performance.

We have tried to improve repetition results by linear low-pass filtering (according to the bandwidth estimated by Nadeu et al.), but reconstruction errors slightly increased. Consequently, we have decided to turn towards nonlinear methods; in particular, we have tried an SVM-approach.

3 An SVM-based reconstruction technique

3.1 SVM

For the reconstruction of the missing frames, we have used a Gaussian Kernel Support Vector Regressor (SVR) [10]. These machines seek to determine a function $f(\mathbf{x})$, that for each data point $\mathbf{x}_i$ verifies $|f(\mathbf{x}_i) - y_i| < \epsilon$. So to speak, the algorithm fits a hosepipe of radius ϵ to the data. The smoothness of the estimated function is controlled by allowing some of the data point to remain outside the pipe.

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

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

where $\mathbf{x}_i$ 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 result of the optimization problem. The SVR 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 SVR, available in [9]. The parameters of the Gaussian kernel have been determined by exploration through a cross-validation procedure.

3.2 The proposed approach

The application scenario entails the following steps: the speech is encoded, packetized and transmitted over an IP network. Once the bit stream reaches the application interface, some parameters are extracted to feed the automatic speech recognizer. In particular, we extract 10 LSPs –Line Spectral Pairs- and the energy of the corresponding frame. Finally, to proceed with the recognition the LSPs coefficients should be transformed into cepstral parameters; however, as this work just asses the reconstruction performance, we have omitted this last stage.

We consider the evolution of each recognition parameter as a time series. A SMV is trained for each parameter to predict the following value from previous ones. It is important to notice that the computational cost of SVM-based reconstruction is not very significant in the ASR context.

Input selection. The feature selection is one of the key points to reach successful results when working with ANN. In this case, we have recovered some prescriptions from the dynamical systems area. In particular, each one of the considered time series is seen as generated by a non-linear dynamical system defined by a (low-dimensional) state-space vector and its evolution through a state space.

As established by the "embedding theorem" [7], it is possible to reconstruct a state space equivalent to the original one. Furthermore, a state-space vector formed by time-delayed samples of the observation (in our case, the speech samples) could be an appropriate choice:

$$\mathbf{s}_n = [s(n), s(n-T), \ldots, s(n-(d-1)T)]^t \tag{2}$$

where $s(n)$ is the time series, d is the dimension of the state-space vector, T is a time delay and t means transpose.

Finally, the reconstructed state-space vector dynamic, $\mathbf{s}_{n+1} = F(\mathbf{s}_n)$, can be learned through either local or global models, which in turn will be polynomial mappings, neural networks, etc.

Considering the reconstructed state-space vector $\mathbf{s}_n$ two questions naturally arise: What should be the embedding dimension of the (reconstructed) state-space vector, d? And what should be the time delay, T? Most of the researchers who have recently proposed non-linear speech predictors have assumed $T = 1$ (following the linear case).

Recently Abarbanel et al. [1] reviewed the state of art concerning the techniques to deal with non-linear deterministic systems. In this paper we propose to apply to our particular problem the analysis techniques described by Abarbanel to determine the time delay and the embedding dimension. It follows a brief summary of these methods.

Average Mutual Information. When seeking the best value for T , the fundamental issue is to establish a right balance between a too small value (samples in the reconstructed state-vector exhibit a lot of common information) and a too large one (samples are independent). Abarbanel et al. suggest the following prescription: choose the value corresponding to the first minimum of the average mutual information

$$I(T) = \sum_{s(n),s(n+T)} P(s(n), s(n+T)) \log_2 \Big[\frac{P(s(n), s(n+T))}{P(s(n))P(s(n+T))} \Big] \qquad (3)$$

where $P(\cdot)$ represents a probability which is estimated through a histogram.

False Nearest Neighbors. Now the issue is to determine the embedding dimension. For that purpose, Abarbanel et al. suggest the false nearest neighbors algorithm which is based on the following reasoning. For any point, we can ask whether its nearest neighbor is there due to the dynamics itself or is instead projected due to a too small reconstructed state-space vector dimension. Thus, the algorithm will compute the percentage of false nearest neighbors (those that disappear when the dimension is increased) for each of the candidate dimensions and will decide that the suitable dimension will be that for which the percentage of false nearest neighbors becomes zero (the dimension is then high enough).

4 Experiments and results

4.1 Input Selection Results

From the Average Mutual Information, $I(T)$, we have obtained that values of T around $T = 23$ could be a good selection for every LSP and around $T = 37$ could be an appropriate choice for the energy. Nevertheless, the first minimum is not well-defined in any case. On the contrary the first valley is quite smooth. Consequently, we consider that other values in a wide neighborhood of these ones can work properly. Even more, for the energy, although the first minimum is reached at $T = 37$, the curve $I(T)$ is extremely smooth and we have decided to use also $T = 23$ for simplicity reasons.

With respect to d, the dimension of the input vector, the results are much more conclusive: $d = 2$ is the best choice in any case.

4.2 Reconstruction Experiments and Results

Database. The database which we have used in our speaker-independent continuous speech recognition experiments is the well-known Resource Management RM1 Database [6], which has a 991 words vocabulary. The speaker-independent training corpus consists of 3,990 sentences pronounced by 109 speakers. 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.

Feature extraction. 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 (see, for example [4]).

Reconstruction results. With the purpose of testing the performance of the SVM-based predictor, we have chosen two subsets from the RM1 database training corpus for 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. Similarly, the SVM validation set provides 36,043 examples.

Table 1. Mean square prediction errors

$Parameters$	$SubstitutionMethod(Repetition)$	$SVMprediction$
$LSP1$	$7.68 \cdot 10^{-5}$	$7.37 \cdot 10^{-5}$
$LSP2$	$12.20 \cdot 10^{-5}$	$11.75 \cdot 10^{-5}$
$LSP3$	$12.58 \cdot 10^{-5}$	$12.10 \cdot 10^{-5}$
$LSP4$	$12.82 \cdot 10^{-5}$	$12.35 \cdot 10^{-5}$
$LSP5$	$15.07 \cdot 10^{-5}$	$14.52 \cdot 10^{-5}$
$LSP6$	$13.86 \cdot 10^{-5}$	$13.42 \cdot 10^{-5}$
$LSP7$	$12.43 \cdot 10^{-5}$	$12.02 \cdot 10^{-5}$
$LSP8$	$13.64 \cdot 10^{-5}$	$13.05 \cdot 10^{-5}$
$LSP9$	$10.99 \cdot 10^{-5}$	$10.49 \cdot 10^{-5}$
$LSP10$	$9.46 \cdot 10^{-5}$	$8.94 \cdot 10^{-5}$
$Energy$	$10.20 \cdot 10^{-3}$	$9.41 \cdot 10^{-3}$

Table 1 shows the results obtained with the conventional and the SVM approaches. As we can see just slight improvements are obtained for every evaluated parameter. Nevertheless, as we have previously indicated, it is expected

that slight reconstruction improvements may lead to significant increments of ASR rates.

5 Conclusions and further work

In this paper we have compared the usual reconstruction method used by speech codecs to circumvent the problems caused by packet losses in IP networks with an SVM-based one aiming at predicting the temporal series described by the codec parameters used by stream-based ASR approaches, obtaining encouraging reconstruction results that would likely improve the recognizer performance.

These preliminary results, however, can be refined exploring different T and d values and from the SVM predictor point of view. We expect improvements using multi-output SVM capable of exploiting the correlations among different LSP parameters corresponding to a particular frame.

6 Acknowledgments

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

References

1. Abarbanel, H.D.I., Frison, T.W. and Tsimring, L.S.: Obtaining Order in a World of Chaos; IEEE Signal Processing Magazine, vol. 15, no. 3, pp. 49-65 (1998)
2. Borella, M. S.: Measurement and Interpretation of Internet Packet Loss, Journal of Communications and Networking, vol. 2, no. 2, pp. 93-102, (2000)
3. 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)
4. Kondoz, A. M.: Digital speech: coding for low bit rate communication systems, Ed. John Wiley & Sons, (1996)
5. Nadeu, C, Pachès-Leal, P. and Juuang, B.-H.: Filtering the time sequences of spectral parameters for speech recognition, Speech Communication 22, pp. 315-322 (1997)
6. National Institute of Standards and Technology (NIST) (distributor): The Resource Management corpus part 1 (RM1) (1992)
7. Ott, E.: Chaos in Dynamical Systems. Cambridge: Cambridge University Press (1993)
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.