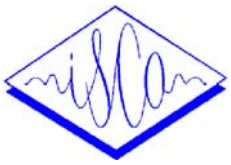


ITRW on Non-Linear Speech Processing (NOLISP 03)
Le Croisic, France
May 20-23, 2003

Some Experiments on Speaker-Independent Isolated Digit Recognition using SVM classifiers

Ramón Fernández-Lorenzana, Fernando Pérez-Cruz, José Miguel García-Cabellos, Carmen Peláez-Moreno, Ascensión Gallardo-Antolín, Fernando Díaz-de-María

Signal Theory and Communications Department, EPS-Universidad Carlos III de Madrid, Spain

1. INTRODUCTION

Hidden Markov Models (HMMs) are, undoubtedly, the most employed core technique for Automatic Speech Recognition (ASR). During the last decades, the research in HMMs for ASR have brought about significant advances and, consequently, the HMMs are currently accurately tuned for this application. Nevertheless, we are still far from achieving high-performance speech recognition-based interfaces.

Some alternative approaches, most of them based on Artificial Neural Networks (ANNs), were proposed during the last decade ([13], [10], [17], [2] and [3] are some examples). Some of them faced the ASR problem using predictive ANNs while others proposed hybrid (HMM-ANN) approaches. Nowadays, however, the preponderance of HMMs is a fact.

Speech recognition is essentially a problem of pattern classification, but the high dimensionality of the sequences of speech feature vectors has prevented researchers to propose a straightforward classification scheme for ASR. Support Vector Machines (SVMs) are state-of-the-art tools for linear and nonlinear knowledge discovery [14], [18]. Being based on the maximum margin classifier, which can be regarded as the common sense solution, the SVM is able to outperform classical classifiers in the presence of high dimensional data even when working with nonlinear machines. The SVM "philosophy" basically states that the only available information for constructing the classifier are the training samples. Therefore, in those applications in which a priori knowledge or structure is known, the SVM might not be as powerful as other machine learning techniques which can benefit form this information. Some work has been done in this direction [7], but still there are open issues that need to be addressed.

Some researchers have already proposed different approaches to speech recognition aiming at taking advantage of this type of classifiers. Among these, [6], [8] and [16] use different approaches to perform the recognition of short duration units, like isolated phoneme and letter classification. In [6], the authors perform a length adaptation based on the triphone model approach. In [8] and [16], a normalizing kernel is used to perform the adaptation. Both cases show the superior discrimination ability of the SVMs. Moreover, in [8], an hybrid approach based on HMMs has been proposed and tested in a CSR (Continuous Speech Recognition) task.

In this paper we propose to use SVMs for speaker-independent isolated digit recognition by plain classification. For this purpose, we use an standard MFCC

parameterization that has been time-adapted to the fixed-input dimension required by SVMs.

The paper is organized as follows.

2. SVM-BASED ASR SYSTEM

In this section we briefly describe the two main blocks of the SVM-based ASR system: the fixed-dimension parameterization module and the SVM classifier.

2.1. Feature extraction and normalization

Since the speech signal varies for each sound and it is not stationary, speech analysis must be performed on short windowed segments, in which the signal is assumed to be quasy-stationary. Typically, a speech signal is divided into a number of overlapping temporal windows (usually the Hamming window is employed) and a speech feature vector is computed to represent each of these time frames. The size of the analysis window, w_s, is usually of 20-30 ms. The frame period, f_p, (the interval between two consecutive feature vectors) is set to a value between 10 and 15 ms. The selected values for these parameters (w_s and f_p) in our particular approach will be discussed later.

With respect to the feature vectors themselves, for each analysis window, twelve Mel-Frequency Cepstral Coefficients (MFCC) are obtained using a mel-scaled filterbank with 40 channels. Then, the log-energy, the twelve delta-cepstral coefficients and the delta-log energy are appended, making a total vector dimension of 26.

Typically, the values of w_s and f_p are kept constant for all utterances and each utterance presents a different time duration. Consequently, the speech analysis generates sequences of feature vectors of variable lengths. However, SVM classifiers require a fixed-dimension input. An alternative for solving this problem consists of adjusting the value of the frame period as a function of the number of samples of each utterance. In this way, it shall be possible to construct sequences of feature vectors of the same length. Following this idea, we have tested two different alternatives, either using variable or fixed analysis window size. In next paragraphs, we briefly describe these procedures.

Variable window size

As in the traditional parameterization procedures, in this case, the window size is chosen to be proportional to the frame period (i.e. $w_s = K\ f_p$), with K (the overlapping factor) being constant for all utterances. Note that K determines the degree of overlap between adjacent analysis windows. Nevertheless, we select first the value of w_s for each utterance in order to obtain the same number of feature vectors for every one. Next, f_p is computed as w_s / K. Figure 1a) illustrates this procedure.

In this way, we are able to provide the SVM with a fixed input vector dimension. However, when the value w_s is too large, the analysed speech segment does not meet the required stationarity properties and an averaged result is obtained. Therefore, we are missing some relevant details of the speech signal.

Fixed window size

In this case, the value of the window size is fixed "a priori" and kept constant for all the utterances. Our aim is working with an analysis window length (25 ms) more consistent with the hypothesis of quasi-stationarity, avoiding to same extent averaged results. However, for obtaining a fixed input vector dimension, we need to dynamically select the frame period (or the overlapping factor, *K*) for each speech utterance. Figure 1 b) shows an extreme example in which analysis windows are not longer overlapped. Therefore, again, some information is missing for long speech utterances.

It is important to notice, however, that through delta parameters (which consider two frames back and forward) some information coming from beyond the actual analysis window is included in the feature vectors.

2.2. SVM training and classification

The SVM given a labelled training data set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ ($\mathbf{x}_i \in \Re^d$ and $y_i \in \{\pm 1\}$, where $\mathbf{x}_i$ is the input vector and y_i is its corresponding label), solves:

$$\min_{\mathbf{w},b,x_i} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} x_i \right\}$$

subject to

$$y_i\left(\mathbf{w}^T \mathrm{f}(\mathbf{x}_i) + b\right) \geq e - x_i$$

$$x_i \geq 0$$

Where $\mathbf{w}$ and b define the linear classifier in the feature space and $\mathrm{f}(\cdot)$ is the non-linear transformation to the feature space ($\mathbf{x}_i \in \Re^d \rightarrow \mathrm{f}(\mathbf{x}_i) \in \Re^H$, $d \pounds H$), unless $\mathrm{f}(\mathbf{x}) = \mathbf{x}$ the solution in the input space will be nonlinear. The SVM minimizes the norm of $\mathbf{w}$ subject to correct classification of all the samples (for every $\xi_i = 0$), in the case any samples can no be correctly classify their corresponding slack variable ξ_i will become nonzero and will be penalised in the objective function. The SVM is usually solved introducing the restrictions in the minimizing functional using Lagrange multipliers, leading to the maximization of the Wolfe dual

$$L_d = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_{ij} a_i a_j \mathrm{f}^{?}(\mathbf{x}_i)\mathrm{f}(\mathbf{x}_j)$$

with respect to a_i and subject to $\sum_{i=1}^{n} a_i = 0$ and $0 \leq a_i \leq C$. This procedure can be solved using quadratic programming (QP) schemes. To solve Wolfe dual, we do not need to know the nonlinear mapping $\mathrm{f}(\cdot)$, only its Reproducing Kernel in Hilbert Space (RKHS) $\kappa(\mathbf{x}_i,\mathbf{x}_j) = \mathrm{f}^T(\mathbf{x}_i)\mathrm{f}(\mathbf{x}_j)$. The value of $\mathbf{w}$ and b can be recovered from the Lagrange multipliers a_i, that are associated with the first linear restriction in the SVM formulation.

For ASR, the input vectors, $\mathbf{x}_i$, will be the concatenation of a fixed number (fixed input dimension) of feature vectors. The class labels 0-9, will have to be modified to be expressed as a binary classification problem. There are basically two approaches to solve binary classification problem with the standard SVM. First, by comparing each class against all the rest (1-vs-all) and each class against all the other classes separately (1-vs-1) [1]. The decisions are usually taken using a majority vote. Also a multiclass SVM has been proposed [19], but is much harder to train and its results are not necessarily better than those of the other approaches. It is still unclear which is the most powerful technique and our experience with 1-vs-all techniques suggest that it is an advisable approach for this problem.

3. EXPERIMENTAL RESULTS

3.1. Baseline System and Database

The baseline is an isolated-word, speaker-independent HMM-based ASR system developed using the HTK package [20]. 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.

We use a database consisting of 72 speakers and 11 utterances per speaker for the 10 Spanish digits. This database was recorded at 8 kHz in clean conditions. Since the database is limited to achieve reliable speaker-independent results, we have used a 9-fold cross validation to artificially extend it. Specifically, we have split each database into 9 balanced groups; 8 of them for training and the remaining one for testing, averaging the results afterwards. In summary, we use a total of 7,920 words for testing our systems.

For the baseline experiment with the HMM classifier, an analysis Hamming windows of 30 ms long was used and the feature vectors (consisting of 12 MFCC, the log-energy, 12 delta-MFCC and the delta-log energy) were extracted once every 10 ms. In this case, both, the window size and the frame period were kept constant for parameterizing all utterances. The average recognition rate achieved by the HMM system was 99,67%. In other terms, only 25 errors over the 7,920 tested words.

3.2. Experiments and Results

Table 1 shows the word recognition rates achieved using both variable and fixed window size and SVM classifiers (always 1 vs. all) as a function of the fixed number of parameters considered as the input for SVM.

As it can be observed, the best result, a word recognition rate of 98,38 %, is obtained for variable window size using 13 feature vectors per utterance. Compared to HMM-based system (99,67 %), the SVM-based system performs slightly worse. Nevertheless, it should be taken into account that the reported results are only preliminary explorations and there are still several issues to be improved and alternatives to be investigated.

4. CONCLUSIONS AND FURTHER WORK

Although these preliminary results are lower that those obtained using HMMs, are not discouraging at all, for two main reasons:

- HMM-based systems has been tuned during the last three decades for this task and, even more encouraging

- Any of the fixed-dimension input vector proposed is missing relevant information of the speech signal, either because uses large analysis windows which lead to averaged results or because some parts of the signal are ignored.

Taking into account the last considerations, we believe that these results are very appealing.

Obviously, though the presented methods need to be refined, it is expected that substantial improvements are achieved through a smarter parameterization procedure. We will work on a more elaborated procedure to achieve the fixed-dimension input. Since the human hearing is relatively insensitive to slowly varying stimuli [9], we propose to focus the spectral sampling on the time instants corresponding the sharpest transitions in the spectral domain. Specifically, we propose to distribute the sampling instants in each utterance according to the derivative of the spectral features. We will start, however, obtaining a non-uniform distribution of sampling instants provided by the internal states of an HMM and a Viterbi decoder. This non-uniform time sampling will give us a first impression about the potentiality of the method.

On the other hand, we expect to extend the SVM framework for ASR along two directions: string kernels and kernel target alignment. The first one, which has been used with success for protein [11] and text [12] classification, could be straight forward extended to speech processing if we were able define a similarity measure for voice utterances. The second approach is a subtle transformation of the kernel matrix to tune its entries to the labels for the given problem [5], [4], significantly improving the performance of the obtained machine. This last approach is mainly described for *transductive learning* [18] in which the whole test set needs to be known a priori, which is its severest limitation.

5. REFERENCES

[1] Allwein, E. L., Schapire, R. E., and Singer, Y, "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers," Journal of Machine Learning Research, 1:113-141. 2000.

[2] Bengio, Yoshua, "Neural networks for speech and sequence recognition", London International Thomson Computer Press , 1995.

[3] Bourlard, Hervé A. and Morgan, Nelson, "Connectionist speech recognition: a hybrid approach", Boston: Kluwer Academic, 1994.

[4] Bousquet, O., and Herrmann, D. J. L., "On the Complexity of Learning the Kernel Matrix," in Advances in Neural Information Processing Systems 15, Editors S. Becker and S. Thrun and K. Obermayer, 2002.

[5] Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J., "On kernel Target Alignment," in Advances in Neural Information Processing Systems 14, Editors T. G. Dietterich and S. Becker and Z. Ghahramani, 2001.

[6] Clarkson, P.; Moreno, P.J, "On the use of support vector machines for phonetic classification", IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2 , pp.585 –588, 1999.

[7] DeCoste, D., and Schoelkopf , B., "Training Invariant Support Vector Machines," Machine Learning, 46. 2002.

[8] Ganapathiraju, A., "Support vector machines for speech recognition" PhD Thesis, Mississipi State University, 2002.

[9] Hermansky, H., Morgan, N., "RASTA processing of speech", "IEEE Trans. On Speech and Audio Processing", vol. 2, no. 4, pp. 587-589, Oct. 1994.

[10] Iso, K. and Watanabe, T., "Speaker-Independent Word Recognition using a Neural Prediction Model", Proc. ICASSP-90, pp. 441-444; Albuquerque, New México, USA, 1990.

[11] Leslie. C., Eskin, E., Weston, J., and Noble, W. S., "Mismatch String Kernels for SVM Protein Classification," in Advances in Neural Information Processing Systems 15, Editors S. Becker and S. Thrun and K. Obermayer, 2002.

[12] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C., "Text Classification using String Kernels," Journal of Machine Learning Research, 2:419-444, 2002.

[13] Sakoe, H., Isotani, R., Yoshida, K., Iso, K., Watanabe, T., "Speaker-Independent Word Recognition using Dynamic Programming Neural Networks"; Proc. ICASSP-89, pp. 29-32; Glasgow, Scotland; 1989.

[14] Schölkopf, B. and Smola, A., "Learning with kernels", M.I.T. Press, 2001.

[15] Smith, N. and Niranjan, M., "Data-dependent Kernels in SVM Classification of Speech Patterns" International Conference on Spoken Language Processing, vol. 1, pp.77-80, 2002.

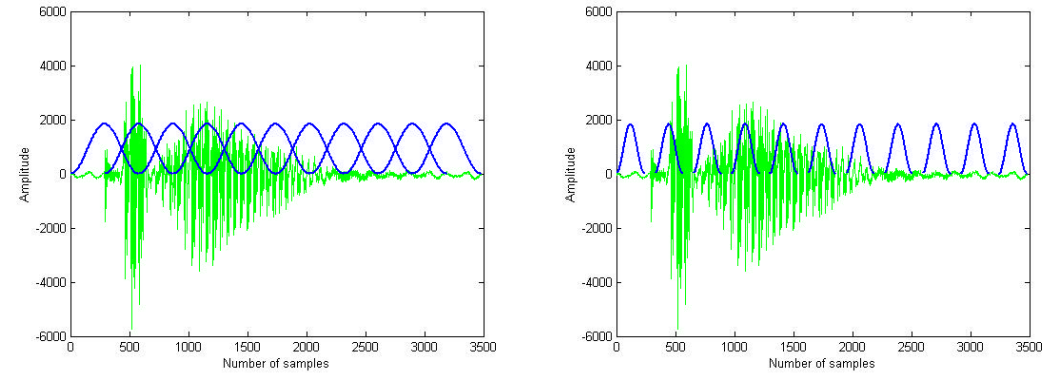
[16] Smith, N.D., Gales, M.J.F., "Using SVMs and discriminative models for speech recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[17] Tebelskis, J., Waibel A., Petek, B., and Schmidbauer, O., "Continuous Speech Recognition using Predictive Neural Networks", Proc. ICASSP-91, pp. 61-64; Toronto, Canada; 1991.

[18] Vapnik, V., "Statistical Learning Theory", Wiley, 1998.

[19] Weston, J., and Watkins, C., "Multi-Class Support Vector Machines," Technical report CSD-TR-98-04. Department of Computer Science, Royal Holloway, University of London, 1998.

[20] Young, S. et al., "HTK-Hidden Markov Model Toolkit (ver 2.1)", Cambridge University, 1995.



a) b)

Figure 1. Two different approaches for normalizing the lenght of feature sequences: variable (figure 1a) and fixed (figure 1b) window size.

Number of feature vectors per utterance	Variable window size Word Recognition Rate (%)	Fixed window size Word Recognition Rate (%)
10	97,99 %	97,01 %
11	97,99 %	96,93 %
12	98,09 %	97,03 %
13	98,38 %	97,03 %
14	98,01 %	96,97 %
15	98,01 %	96,94 %

Table 1. Recognition results using a SVM classifier and the two proposed time-normalized parameterizations: variable or fixed window size.