Classification of Spanish Vowels and Digits using Neural Networks.

JOSE BRITO and WLADIMIR RODRIGUEZ
Postgrado en Computación, Facultad de Ingeniería
Universidad de Los Andes
Núcleo La Hechicera, Edif. B, 3º piso, Ala sur
5101 Mérida,
VENEZUELA
http://www.pgcomp.ula.ve

Abstract: - In this paper we describe the use of Multilayer Perceptron Array for learning and classifying speech signals, using characteristic vectors of reconstructed dynamics. First, we consider the phonatory system as a black box, where the only available data is its output: the speech signal. This is a way of accessing underlying dynamics, and is the starting point for two kinds of experiments: classification of vowels and digits, with Venezuelan Spanish voices. Results verify positively that characteristics vectors extracted from underlying dynamics hold discriminative power for distinguishing between classes of speech signals. Besides, neural networks are able to generalize using this kind of data.

Key words: Speech signals, Multilayer Perceptron, Reconstructed State Space, Classification .

1 Introduction.

Lately, it has been a surge in the interest of the parameterization on speech signal based on the reconstruction of the phonatory system's dynamics [5, 6]. In principle, a pattern recognition problem must be solved using a qualitative approach over the underlying physical systems profile [10, 11], in this case, the phonations. On the contrary, the conventional analysis techniques are based on the hypothesis that the verbal signal is linear, although there are many objections to this assumption. For example, in the popular source-filter model of speech generation, the excitation source is the turbulence developed in the vocal track itself, so in this case the natural source does not match the model. Besides, a linear model will have a difficult time adjusting to high variability in the signal. Therefore, in the end, these models are only an approximation of the phonatory system. The approach we present here recurs to a simple Multilayer Perceptron Array (MPA) for learning and classifying speech signals, using characteristic vectors of reconstructed state space.

Two kinds of experiment are performed: vowels and digits. Signals in both trials are extracted from SpeechDat [4] database of Venezuelan Spanish utterances. Therefore, this is the first study, which employs nonlinear techniques with Venezuelan voices. In each experiment, two corpora are defined: C_E and C_P, consisting of signals for training the MPA, and testing it, respectively.

2 Reconstructed state space

In the case of nonlinear systems, with incomplete data, the extraction of new information from data is more difficult than in the linear case [2]. If the system is very complex (ie. phonatory system), but only one of its properties (ie., verbal signal) is available by a sensor, the traditional analysis procedures will be very limited. As an alternative, the reconstructed state space allows to recover the nonlinear system's dynamics from only one time series [1]. Basically, in this space some geometric structures called attractors are build by the trajectories. Naturally, the reconstructed space is not completely equivalent to the internal system's dynamics, but under some theoretic restrictions, the topology of the dynamics is preserved. This allows that the conclusion obtained from the reconstructed dynamics will be valid for the real and inaccessible internal dynamics (black box) [1, 7, 9]. Also, it would help on the detections of the time series' structures that could pass unnoticed.

It follows the description on how to obtain the reconstructed state space. Consider a set of samples uniformly spaced of one verbal signal S_v. The reconstructed state space is a multidimensional representation of the signal against delayed versions of itself (subseries). In more formal terms, the reconstructed space state is formed by the definition of the vectors $\mathbf{V}_n$ en $\Re^m$:

$$\mathbf{V}_n = \{S_v[n], S_v[n+\tau], ..., S_v[n+(m-1)\tau]\} \quad (1)$$

or

$$\mathbf{V}_n = \{S_v[n], S_v[n-\tau], ..., S_v[n-(m-1)\tau]\} \quad (2)$$

where $S_v[i]$ is the signal value in time i (a sample). In turn, m and τ are fundamental reconstruction parameters known as embedding dimension and lag, respectively. The Takens' theorem [9], which associate the Reconstructed State Space with the real internal systems' dynamics, express that given sufficient m and τ, the real dynamics and the Reconstructed State Space are topological equivalent. Preliminary tests with differential entropy method [3] resulted in low values for m and τ, over the corpus of vowels. So, we set $m = 2$ and $\tau = 3$ in subsequent experiments. Figure 1 shows the RSS for an arbitrary vowel, with a grid defining 100 blocks over the plane. On the other side, digits constitute very complex signals because of their superior phonetic richness, and consequently, simple space representation is out of the question. A somewhat different approach, discussed in next section, will be used for digits.

Note that each axis in the figure corresponds to [-1, +1] interval, which is achieved by means of normalizing the speech signal: every sample is divided by $max(abs(S_v))$. This trivial, but important step allows the blocks B_i to be of fixed dimensions. Strictly, each block is a square, and its area is $4 / r$ (dimensionless), where r is the total of blocks over the plane. In figure 1, $r = 100$, and so each block's area is 0.04.

3 Feature Extraction

For the vowels, the characteristics vector V_C^V is given by the spatial density for each block B_i($1 <= i <= r$). Then, V_C^V has, in principle, r elements. In each B_i, the spatial density is compute as follows:

$$spatialDensity(B_i) = \frac{|B_i|}{|S_v|} \quad (3)$$

where, $| B_i |$ is the numbers of points of the attractor delimited by B_i and $|S_v|$ total of samples.

Also, to give robustness to the classification, V_C^V is extended with the resulting r elements of the previous analysis over the time series $S_v^d = S_v[i + 1] - S_v[i]$. This way, the signal's variation speed is included to the vector, by means of an approximation of the first differences.

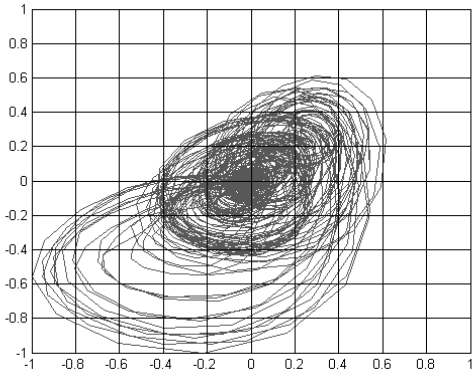

Fig. 1: RSS for a Venezuelan Spanish vowel

In the digits case, the described analysis is applied over superposed plots, without the r elements of S_v^d. Here, a plot is a sequence of samples, in other words, a subsequence of S_v. For a given signal, the plots always have the same size, but between signals, the plot's size could change. Exactly they are named proportional because their size is a proportion of the signal. The signal is divided in np same sizes segments , np is even. Where $LSeg$ is the length of each one of the np segments ($LSeg = Length(S_v)/np$), then the plot's length LTr is 2 x $LSeg$. The first plot starts with the first sample of the signal. After that, each new plot starts at the middle sample of the previous one with an extension of LTr samples. Therefore, the $np - 1$ plots of a signal start at the sample

$$i \text{ x } LSeg + 1 \ (0 \leq i \leq np - 2)$$

Later, an analysis similar to the one for vowels is applied to each plot. Of this form, the plot's analysis contributes to the digits characteristic 's vector, V_C^D, with $(np - 1)$ x r elements.

4 Multilayer Perceptron Array

Be n the amount of categories to recognize. Then, for each experiment, the classifier consists of an array $[R_1R_2...R_n]$ of n multilayer perceptrons neural nets R_i. Thus, a neural net is associated with each category. Once the corpus C_E is defined, we can start the training session. Basically, each R_i is trained with all the signals j ($1 \leq j \leq cardinality\ (C_E)$) of C_E. To all the training entries for which it can be verified that category(j) = category(i), the output will be 1; else the output is 0.

Later, when classifying, the input signal is

characterized and the resulting vector is administered to each one of the *n* neural nets. The neural net with the highest output determines the category for the signal.

The neural nets used for the classifier have three layers. The number of input neurons will depend on the size of the characteristics vector, V_C^V or V_C^D. After that, if that vector has *p* components, then we will have *p* input neurons. For example, if the state space is partitioned in 100 blocks, and the density of each one is calculated, V_C^D, this will implied that the input layer will have 200 elements. On the other hand, in the hidden layer will have 5 neurons, and the output layer one neuron. The activation functions are logarithmic sigmoids, with the exception of the output neuron that have a linear transfer function. Finally, the Levenberg-Marquardt algorithm is used for training the neural net. This is an advanced algorithm for non-linear optimization, and normally converges to the minimum error faster that the Back-Propagation one, although it has a big memory requirement.

5 Results

For each experiment, to be done, a MPA was build. We set $r = 100$, and for digits we set *np* in 6. In order to verify the classifiers, we gave as input the signal in the training corpora, obtaining a 100% classification rate. Speaker dependent tests gave, in average, recognition's rates of about 92% and 65.5%. In the other hand, with speaker independent tests, we obtain, for the vowels; an average of 55%, and 66% for the digits.

	a	**e**	**i**	**o**	**u**	**%**
a	8	2	2	8	0	40.00
e	0	8	9	3	0	40.00
i	0	4	12	2	0	60.00
o	2	2	0	14	2	70.00
u	0	1	3	3	13	65.00
						55.00

Table 1: Confusion matrix for speaker-independent vowels.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**%**
0	11	0	0	0	0	2	0	7	0	0	55.00
1	0	12	2	4	0	0	0	0	1	1	60.00
2	0	0	13	3	1	0	0	0	0	3	65.00
3	2	0	2	14	0	1	0	1	0	0	70.00
4	1	0	0	0	17	0	0	2	0	0	85.00
5	1	0	0	0	0	11	0	8	0	0	55.00
6	2	1	0	4	1	0	9	1	0	2	45.00
7	3	0	0	0	1	2	0	14	0	0	70.00
8	0	0	1	0	1	0	0	0	18	0	90.00
9	0	3	0	2	0	1	0	1	0	13	65.00
											66.00

Table 2. Confusion matrix for speaker-independent digits

We can see that the main diagonal of the confusion matrixes confirm the tendency of the MPAs to correctly classify the input signal. In the case of the vowels, the variability between the speaker's independent signals deteriorates the recognition accuracy. It is interesting, that this does not happen in the case of the digits, maybe because these signals include more information than the vowels, and the MPA is able to capture it.

The characterization of the attractor's density in the state space has been considered by some studies, to classify vowels signals. For example, in [10] a Bayesian classifier is used, with an average accuracy of 34.49% for the English vowels. In [5,6] a fuzzy information space is used, with 100% accuracy but only with a six signals corpus. Comparing with frequency domain techniques, we have the work of Maldonado [4], which used a corpus about Venezuelan Spanish digits, with recognition rates above 90%. But, this work used already establishes analytic approximations, like cepstral coefficients and hidden markov models.

6 Conclusions

Considering that this analysis is completely time-domain based, recognition rates are fairly good. However, further research is needed for determining effect of a higher (ie., $m > 2$) dimensional analysis. When more than two dimensions are used,

characterization becomes difficult. Except for MPA training, the exposed techniques do not require considerable computational resources. Then, a question to answer is if computationally intensive, high dimensional analysis, results worthy, taking into account the current accuracy of frequency domain techniques. This kind of investigation is needed because learning of more phonemes will certainly ask for more attractor data.

References

[1] H. Abarbanel, R. Brown, J. Sidorowich, y L. Tsimring, "The analysis of observed chaotic data in physical systems", *Reviews of Modern Physics*, vol. 65, No. 4, 1993.

[2] E. Bradley, "Time series analysis", in *Intelligent Data Analysis: An Introduction*, Springer, 1999.

[3] T. Gautama, D. Mandic, y M. Van Hulle, "A differential entropy based method for determining the optimal embedding parameters of a signal", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2003.

[4] J. L. Maldonado, *Tratamiento y reconocimiento automático de señales de la voz venezolana*, Disertación doctoral, Universidad de Los Andes, 2003.

[5] W. Rodríguez, H.-N. Teodorescu, F. Grigoras, A. Kandel y H. Bunke, "A fuzzy information space approach to speech signal non-linear analysis", *International Journal of Intelligent Systems*, vol. 15, No. 4, pp. 343-363, 2000.

[6] W. Rodriguez, "Similarity of Dynamical Systems", Ph.D. Thesis, University of South Florida, 1998.

[7] T. Sauer, J. A. Yorke, y M. Casdagli, "Embedology", *Journal of Statistical Physics*, vol. 65, pp. 579-616, 1991.

[8] Shepherd, A. J. *Second-Order Methods for Neural Networks*, Springer-Verlag, 1997.

[9] F. Takens, ‘’Detecting strange attractors in turbulence’’, *Dynamical Systems and Turbulence*, Warwick, 1980.

[10] J. Ye, M. T. Johnson, y R. J. Povinelli, "Phoneme Classification using Naive Bayes Classifier in Reconstructed Phase Space", *10th IEEE Digital Signal Processing Workshop*, 2002.

[11] F. Zhao, "Extracting and Representing Qualitative Behaviors of Complex Systems in Phase Space", *Artificial Intelligence*, vol. 69, pp. 51-92, 1994.