New Method for Classification of ASR Word Hypotheses Using Prosodic Cues

Enrique M. Albornoz[†,‡], Diego H. Milone[†,‡], Hugo L. Rufiner[†,‡], Ramón Lopéz-Cózar[§]

[†]*Centro de Investigación en Señales, Sistemas e INteligencia Computacional (SINC(i))*
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral
sinc@fich.unl.edu.ar
[‡]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)*
[§]*ETSI Informática y de Telecomunicación, Universidad de Granada, Granada (España)*
rlopezc@ugr.es

***Abstract*— In this work, we propose a novel resampling method based on word lattice information and we use prosodic cues with support vector machines for classification. The idea is to consider word recognition as a two-class classification problem, which considers the word hypotheses in the lattice of a standard recognizer either as True or False employing prosodic information. The technique developed in this paper was applied to set of words extracted from a continuous speech database. Our experimental results show that the method allows obtaining average word hypotheses recognition rate of 82%.**

1 Introduction

Over the last years, prosodic information has become a very interesting line of research. A lot of efforts have been made to model and incorporate it in Automatic Speech Recognition (ASR) systems. In doing so, two important issues must be considered. On the one hand, extracting and modelling the prosodic elements to be employed, whilst on the other, finding the best way to incorporate them in an ASR system. A number of papers can be found in the literature addressing these issues.

For example, [7] proposed to use a combination of prosodic features and accentuation to model Spanish words. A prosodic binary classifier for syllable stress that is used with ToBI (Tones and Break Indices) [10] information to evaluate the ASR hypotheses is defined in [2]. In [11], prosodic information is used to train a small set of Hidden Markov Models (HMM) in order to segment prosodic units in Hungarian language. A method where prosodic features in syllables are categorised in 16 classes using vector quantisation, and words are defined as a concatenation of these classes is proposed in [6]. In [12], the supra-segmental features of speech are modeled with prosody in a traditional HMM framework. This method is designed for fixed-stress languages where a segmentation for syntactically linked word groups is done. A prosodic model for Spanish word classification was proposed in [1]. It uses the orthographic rules of Spanish to do groups of words depending on the separation of the syllables. In every word, prosodic information is compared among syllables in order to obtain a code of the relative magnitude measured in each one.

However, some problems arise with the ASR system when prosodic analysis is in the level of syllables. For example, confusions do not only appear among words with the same number of syllables, and for this reason the information from orthographic rules is not so useful. Another problem is that the recogniser usually makes mistakes for some particular words. Using word nets, an additional problem is that nets do not always have the true hypothesis in every speech segment.

In this paper, we present a method to address errors of the acoustic models typically employed in a standard HMM-based speech recogniser. We propose to develop word classifiers to identify the incorrect hypotheses in problematic speech segments. Moreover, we propose an original alternative to tackle the problem of choosing the proper data to train these classifiers. On the other hand, the incorporation of this information in an ASR system will be considered in future works.

The remainder of the paper is organised as follows: in Section 2 the proposed method is presented, where it is explained a new resampling methodology for a speech corpus and how to use it in order to classify word hypotheses; Section 3 introduces the features extraction process and discusses an experiment that, for each word, explores different configurations, features vectors and classifiers; then, the Section 3 presents a second experiment which uses the best configurations and test data; finally, Section 4 presents conclusions and discusses possibilities for future work.

2 The proposed method

Usually, the first step in a simple word classifier is to extract acoustic segments and label them according to the correponding word in the utterance. From these segments, different features are computed and selected to compose the inputs for the classifier. In this way, after the training phase, the classifier would be able to predict one word from a set of features that it has never seen before.

State-of-the-art HMM-based ASR systems [9] may have good performance in appropriate conditions, but sometimes have problems with particular words, for example, due to the accents of the speakers. Thus, the focus of our method is the re-analysis of these problematic segments.

The proposed methodology is based on the analysis of the recognition hypothesis space provided by an ASR system when it recognises an utterance. It requires creating an HMM-based ASR system in the standard way and generates N-Best word lattices for all training utterances. These lattices are used to build a lattice corpus by resampling, which is used to train word classifiers with additional features. The resampling process for the classifier is explained in the next sections.

2. 1 N-best hypotheses resampling

The speech signals used in the experiments were taken from the Albayzin corpus, a Spanish continuous speech database, developed by five Spanish universities [8]. The corpus utterances were spoken by twelve people, six females and six males. In the experiments, we have used 4400 utterances corresponding to the training set in the corpus.

In order to create a standard HMM-based ASR system, we used the *Hidden Markov Model Toolkit* (HTK) [13]. The classic Mel-frequency cepstral coefficients (MFCC) parameterisation was calculated using a Hamming windows of 25 ms with a 10 ms frame shift. The first 12 MFCCs and the energy plus their first and second derivatives were extracted for phoneme modelling, and a bi-gram language model was generated.

Then, we created an N-best list of hypotheses ($N = 10$) for every training utterance. Acoustic segments were extracted from the utterances using the information about the alignments of word hypotheses. For each word in the utterance, the word hypotheses were inserted either in a set called True or in a set called False, depending on the correspondence between the hypotheses and the orthographic transcription of the utterance. For example, Figure 1 shows a word lattice for a speech segment where some hypotheses match with the transcription. In this example, there are three True hypotheses for the word *dime*, one False hypothesis for the word *casa* and one False hypothesis for the word *grande*.

We have defined two rules in order to balance the sets of True and False hypotheses obtained.

- All the repeated True hypotheses are discarded to avoid redundancy.

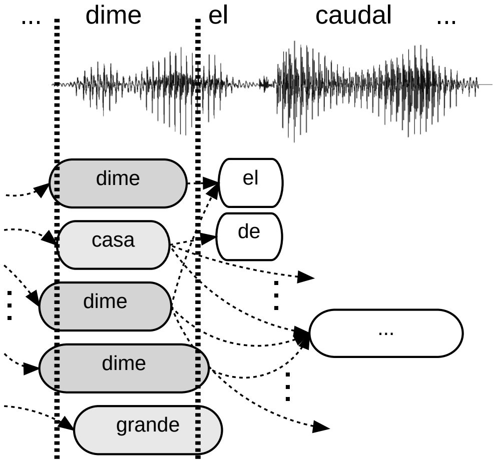

Figure 1: N-best instance for a speech segment.

- All False hypotheses are kept because little redundancy is found in this set.

The second rule allows considering more varied True hypotheses. The data in the two sets is resampled to balance the size of the sets. To do so, we consider the following rules:

1. If $\mathit{count}(\mathit{True}) > \mathit{count}(\mathit{False}) \Rightarrow$ the True set is defined by simple random sampling without replacement of True data.

2. If $\mathit{count}(\mathit{True}) < \mathit{count}(\mathit{False}) \Rightarrow$ the False set is defined as the unreplicated False data plus simple random sampling without replacement of these data.

2. 2 Feature extraction and classification models

As discussed in Section 1, prosodic features such as F_0 and energy have been extensively used for ASR [2, 11, 6]. Many prosodic parameters can be extracted from these features, for example, mean, minimum, maximum and slopes. In the next section, the chosen parameters are explained.

As our method requires a binary classifier and one of the two sets of hypotheses is very populated, we have used support vector machines (SVMs). A SVM is a supervised learning method widely used for pattern classification, which has theoretically good generalisation capabilities. Its aim is to find a hyperplane able to separate input patterns in a sufficiently high dimensional space [3]. In the experiments we have used the *LIBSVM* library [5] to process the patterns obtained from the prosodic parameters. The proposed method implements a *one-against-all* classification scheme where *one* represents the true hypotheses (given that there are many and diverse false hypotheses). Therefore, the classifier should fit the frontier region for the True class and the remaining space should be for the False class. Following this approach, the classifier can deal with

Table 1: Best classification results (in %) for training data and raw features.

Word	*42*	*32*	*26*	*21*	*16*	*14*	*12*	*10*	*8*	*6*	*4*	*2*
CABO	61.11	63.89	63.89	63.89	63.89	63.89	70.83	76.39	77.78	**79.17**	76.39	62.50
CAUDAL	76.52	80.30	80.30	80.68	80.68	79.55	80.68	**85.61**	84.85	84.47	80.30	77.65
DESEMBOCA	80.38	80.38	80.38	80.38	80.38	80.38	80.38	80.38	80.38	**84.21**	80.38	74.64
DESEMBOCAN	79.79	79.79	79.79	83.94	84.72	84.72	84.72	84.72	84.72	**85.49**	**85.49**	63.73
MENOR	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	**75.76**	74.89
MENOS	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**	**81.63**
NOMBRE	86.75	86.75	87.73	87.83	**87.93**	87.73	87.54	87.63	87.63	86.95	86.75	79.49
NUMERO	84.85	84.85	84.85	84.85	84.85	84.85	84.85	84.85	84.85	84.85	**89.39**	87.88
PASA	79.17	79.17	80.11	80.11	80.30	80.11	79.36	79.36	79.36	79.55	**80.49**	73.48
PASAN	56.22	56.22	56.22	56.22	56.22	56.22	56.22	56.22	57.25	59.33	61.66	**65.80**
TIENE	52.66	52.66	52.66	52.66	52.66	52.66	52.66	**76.86**	75.27	73.27	71.94	65.82
TIENEN	69.08	69.08	69.08	71.60	72.27	72.10	**73.95**	73.45	73.61	70.92	68.57	64.71

Table 2: Best classification results (in %) for training data and normalised features.

Word	*42*	*32*	*26*	*21*	*16*	*14*	*12*	*10*	*8*	*6*	*4*	*2*
CABO	77.78	81.94	77.78	81.94	86.11	86.11	**90.28**	73.61	75.00	75.00	75.00	66.67
CAUDAL	**89.77**	89.02	87.50	85.61	83.33	82.95	84.47	86.36	85.23	81.82	76.14	74.62
DESEMBOCA	84.93	**85.65**	85.41	82.78	84.21	82.78	83.49	81.34	81.82	71.05	65.79	61.24
DESEMBOCAN	80.83	80.05	**81.87**	79.53	79.02	78.50	77.72	75.13	78.24	69.95	62.69	58.55
MENOR	88.31	88.31	**89.18**	85.71	87.01	86.58	85.28	86.15	84.85	83.12	75.32	73.16
MENOS	86.39	**87.07**	85.71	86.39	86.39	85.71	85.71	85.03	82.99	84.35	82.31	73.47
NOMBRE	**88.32**	**88.32**	88.22	87.34	85.87	86.46	83.91	80.77	79.39	73.80	72.42	71.64
NUMERO	**89.39**	86.36	84.85	86.36	80.30	78.79	83.33	84.85	81.82	81.82	81.82	75.76
PASA	84.28	83.90	**84.47**	83.14	81.82	79.73	79.17	77.46	74.43	74.05	69.89	69.32
PASAN	74.61	**76.42**	74.35	75.13	72.80	74.09	75.13	74.61	72.54	68.13	69.43	65.54
TIENE	**78.19**	77.79	75.93	76.46	73.01	73.54	73.40	71.68	69.81	68.35	66.22	63.16
TIENEN	**75.97**	74.62	72.61	72.94	72.61	70.42	71.26	66.22	67.56	64.37	63.53	64.20

word hypotheses not observed in the training, which may correspond to out-of-vocabulary words.

3 Experiments

We first discuss how we have chosen the feature vectors and the best SVM model for each word. Then, we report on the tests that we have carried out using these models, with data partitions not observed in the training.

In this work, twelve of the most confused words were selected according to the ASR errors. These were computed in the N-Best extraction stage. For every word, a training/test partition with the balanced corpus was generated. 80% of the data was randomly selected for training and the remaining 20% was left for test. The experiments were performed using raw data on one hand, and normalised data on the other, in order to compare the relevance of the normalisation step. Each feature dimension was independently normalised in the training stage, using the maximum and the minimum of each dimension. Then, these scale factors were used in the test stage.

We used the Praat toolbox [4] to extract F_0, Energy, F_1, Bandwidth of F_1, F_2 and Bandwidth of F_2 from the recognition hypotheses. Their minimum, mean, maximum, standard deviation, skewness and kurtosis coefficients were also computed to create features vectors (FV) that have 42 features: the mentioned 36 features plus minimum and maximum distance between F_1 and F_2, mean square error between F_1 and F_2, and F_0, F_1 and F_2 slopes.

For each word, the F-Score measure was used to rate the features depending on their discriminative capacity [5]. Given the feature vectors FV_k, this score was computed considering the True instances (N_T) and the False instances (N_F) as follows:

$$F(i) = \frac{\left(\bar{x}_i^{(T)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(F)} - \bar{x}_i\right)^2}{\frac{1}{N_T-1}\sum_{j=1}^{N_T}\left(x_{j,i}^{(T)} - \bar{x}_i^{(T)}\right)^2 + \frac{1}{N_F-1}\sum_{j=1}^{N_F}\left(x_{j,i}^{(F)} - \bar{x}_i^{(F)}\right)^2}$$

where $\bar{x}_i$ is the average of the ith feature, $\bar{x_i}^{(F)}$ and $\bar{x_i}^{(T)}$ are the average False and True instances respectively, and $x_{j,i}$ is the ith feature in the jth instance.

In a first experiment, using the rated features we created 12 different input patterns for each word considering the 2, 4, 6, 8, 10, 12, 14, 16, 21, 26 and 32 most discriminative features on the one hand, and all the features (42) on the other. For each feature set, SVM parameters were explored in order to create the best classification model. Every SVM model used a radial basis function kernel, the accuracy of which was computed using a five-fold cross validation scheme, considering the training data only. As a result we obtained

Table 3: Word hypotheses classification results for test raw data.

Word	Selected subset vector	Accuracy[%]
CABO	6	66.67
CAUDAL	10	74.24
DESEMBOCA	6	89.42
DESEMBOCAN	6	85.42
MENOR	42	77.19
MENOS	42	67.57
NOMBRE	16	90.20
NUMERO	4	75.00
PASA	4	81.06
PASAN	2	56.25
TIENE	10	82.98
TIENEN	12	85.91
Average accuracy		**77.66**

Table 4: Word hypotheses classification results for test normalised data.

Word	Selected subset vector	Accuracy[%]
CABO	12	66.67
CAUDAL	42	84.85
DESEMBOCA	32	89.42
DESEMBOCAN	26	82.29
MENOR	26	91.23
MENOS	32	83.78
NOMBRE	42	85.49
NUMERO	42	81.25
PASA	26	81.82
PASAN	32	77.08
TIENE	42	75.00
TIENEN	42	86.57
Average accuracy		**82.12**

the best parameters for each feature set. Tables 1 and 2 show the classification accuracy for raw and normalised training data. In these tables, the number of selected features is showed in the first row.

In a second experiment, a new SVM model was trained with the whole training data for each word, using the settings that achieved the best accuracy in the first experiment. All SVM models were tested with the aforementioned test partitions. Tables 3 and 4 set out the results obtained. It can be observed that these models achieved good results classifying word hypotheses. The average recognition rate improved when normalisation was applied, but this process required more features. It should be noted, however, that the normalisation process is not very useful for all words, as can be observed in the tables. For example, the classification rate for the word *MENOR* was 77, 19% using raw features and 91, 23% using normalised features, whereas for the word *NOMBRE* it was 90, 20% using raw features and 85, 49% using normalised features. This suggests that the normalisation process could be customised for each word in order to improve the performance.

4 Conclusions and future work

In this work, we have presented an approach aimed to improve the performance of standard ASR systems, which considers word lattices and prosodic cues. In accordance with this method, firstly, word lattices generated by a standard HMM-based speech recogniser are used to extract word hypotheses. Secondly, these hypotheses are the input to single-word classifiers that distinguish between True and False hypotheses considering prosodic information. The experimental results show that the method achieves average word accuracy of 82% when applied to a speech database in Spanish. Although more experimentation is needed, these results are promising in order to get an improvement in the performance of a standard ASR system. Moreover, the method could be applied to any language as it is language-independent because the method does not include any specific Spanish rule.

In future work we will integrate the method in a standard ASR system to increase the probabilities of the true hypotheses in the recognition network. Classifying word hypotheses using prosodic features would allow to process a real ASR problem efficiently. Results indicate that every word should be dealt with a specific model configuration in order to improve the recogniser performance. In addition, we plan to work on an "one-pass" system that, using our method, will take as input the alignments of the hypotheses and will produce the ASR result.

5 Acknowledgements

The authors wish to thank: *Agencia Nacional de Promoción Científica y Tecnológica* and *Universidad Nacional de Litoral* (with PAE 37122, PAE-PICT 00052, CAID IIR4-N14) *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina, *Universidad de Granada from España*, *Programa Erasmus Mundus External Cooperation Window - Lote 16 EADIC* for their support.

References

[1] Enrique M. Albornoz and Diego H. Milone. Construcción de patrones prosódicos para el reconocimiento automático del habla. In *34th JAIIO*, pages 225–236, Rosario, Argentina, September 2005. Simposio ASAI.

[2] S. Ananthakrishnan and S. Narayanan. Improved Speech Recognition using Acoustic and Lexical Correlates of Pitch Accent in a N-Best Rescoring Framework. *Acoustics, Speech and Signal Processing. ICASSP 2007. IEEE International Conference on*, 4:873–876, April 2007.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1 edition, August 2006.

[4] Paul Boersma and David Weenink. Praat: doing phonetics by computer (Version 5.1.32) [computer program], 2010.

[5] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[6] Songfang Huang and Steve Renals. *Machine Learning for Multimodal Interaction*, volume 4892/2008 of *LNCS*, chapter Using Prosodic Features in Language Models for Meetings, pages 192–203. Springer Berlin, 2008.

[7] D. H. Milone and A. J. Rubio. Prosodic and accentual information for automatic speech recognition. *IEEE Trans. on Speech and Audio Proc.*, 11(4):321–333, July 2003.

[8] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Marino, and C. Nadeu. Albayzin speech data base: design of the phonetic corpus. In *Proc. of the 2th European Conf. of Speech Communication and Technology*, pages 175–178, Berlin, September 1993.

[9] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[10] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: a standard for labeling english prosody. In *Proc. of International Conference on Spoken Language Processing ICSLP '92*, pages 867–870, Banff, Alberta, Canada, October 1992.

[11] G. Szaszák and K. Vicsi. *Verbal and Nonverbal Communication Behaviours*, volume 4775/2007 of *LNCS*, chapter Using Prosody in Fixed Stress Languages for Improvement of Speech Recognition, pages 138–149. Springer, Berlin, Heidelberg, 2007.

[12] K. Vicsi and G. Szaszák. Using prosody to improve automatic speech recognition. *Speech Comm.*, 52:413–426, May 2010.

[13] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for HTK Version 3.1).* Cambridge University Engineering Department, England, December 2001.