Including prosodic cues in ASR systems

Diego H. MILONE
Universidad Nacional del Litoral, Universidad Nacional de Entre Ríos
Argentina
d.milone@ieee.org

and

Antonio J. RUBIO
Dept. Electrónica y Tecnología de Computadores
Universidad de Granada, Spain
rubio@ugr.es

ABSTRACT

Several aspects related to production, as well as natural perception of speech, have gradually been incorporated to automatic speech recognition systems. Nevertheless, the set of speech prosodic characteristics has not been used for the time being in an explicit way in the recognition process itself. In this work, an analysis of the prosody's three most important parameters: energy, fundamental frequency and duration, is presented with a method to incorporate this information into automatic speech recognition. Prosodic-accentual features are incorporated in a hidden Markov models recognizer. Their theoretical formulation and experimental setup are presented. Several experiments are developed to show the method behavior in a Spanish continuous speech database. From this understanding and with other database subsets, the overall results provide a word recognition error reduction that would reach more than 30% when prosodic-accentual cues are incorporated.

Keywords: Automatic speech recognition systems, prosodic cues, features extraction, language model, accentual structure estimation.

1. INTRODUCTION

Since the introduction of Hidden Markov Models (HMM) to automatic speech recognition (ASR), this technique has become a "de facto" standard in most laboratories worldwide.

The reasons for this generalization of the use of HMM are mainly two: the first one is the easy estimation of the acoustic models, provided an adequate database is available, and the second reason is related to the good results than such HMM-based systems can achieve.

In practice the size of acoustic models has to be rather small to reduce the number of them to be estimated. The larger the number of acoustic models, the larger the database has to be to maintain a given parameter estimation quality.

Thus, most ASR systems use phone-like units. This is the only possibility of having a small enough number of acoustical models.

The length of these phone-like units is about the length of a phoneme. Therefore, they do not contain much information about prosodic features. Although many exhaustive investigations around ASR are carried out, very few of them are related with including prosodic cues in the recognition process.

The influence of prosody in the human speech recognition process is obvious and some works related to improving Text to Speech Synthesis (TTS) can be found [17]. The results of these kind of works offer many ideas and provide important data about the natural way in which the human being uses the prosody in spoken discourse. Also, prosodic information has been used in speaker identification [18]. In ASR, by simply taking into account of the speaking rate, improvements in recognition rate can be made [3].

For these reasons with face the problem of incorporating prosodic cues to the automatic speech recognition process. There exist much information in a sentence that is not considered

in the standard recognition process with phone-like acoustic units.

2. BACKGROUND

It is generally recognized that including prosodic features in ASR systems would produce some benefits but no concrete solution is normally proposed.

Some projects have incorporated some of the prosodic characteristics to solve only a reduced group of problems related to ASR. We will mention here some of them.

In [12], the authors use pitch to recover some recognition errors in a connected digits task. This is an example of the class of papers that propose a post-recognition incorporation of the prosodic features and not as a part of the recognition system itself. Other instances of this class are [2], [14], [19], [21] and [23], where a N-best recognition output is rescored using prosody.

In other cases, like [20], a pre-segmentation based on prosodic features is carried out to perform recognition on the obtained parts.

There are few cases in which the explicit incorporation of the prosody is carried out during the recognition process, but in these cases the use of this information is rather limited, like in [10] and [22] where prosodic features help in the detection of end of sentence and other boundaries.

Also, some projects related to tonal languages have incorporated pitch to the recogntion process, as in these languages there is a more direct relation between the tonal cadence and the meaning of a word [6], [7], [9], [10], [11].

3. PROSODY

Stress is an aspect closely related to prosody and an important part of this work. In different languages prosodic characteristics are really close to accentuation [5], [1], [8].

From a physical point of view, prosody can be defined as the effect of different combinations of energy, pitch and duration values in the spoken language. These are the characteristic usually considered when speaking about prosody.

From a linguistic point of view, the important issue is how prosody is produced from different abstraction levels in spoken language [16].

We have to face the twofold problem of obtaining the prosodic features and incorporating them to the ASR process.

4. STRESS

Accentuation is a characteristic of the prosody that is very language-dependant. In French the stress is always found in the last syllable, as in Finnish it is always located in the first syllable of a word.

In Spanish, like other languages (English, German or Italian) has a free accentual topology. That is, the stress can be found in any sillable of a word.

Stress is related to supra-segmental prosodic features in Spanish. The vowel of a stressed sillable has greater values for energy, fundamental frequency and duration, although the position of the stress may change from an isolated word to the same word embebed in continuous speech [16], [24].

For Spanish, more than 36% of words can be considered as atonic (non-stressed sillables, 90% of them are monosillabic words), although some of them change their accentual pattern depending of the grammatical function [16]. We have used the information found in [16] as a good starting point. Nevertheless, we made an additional analisys to better knowing the link between stress and prosody in continuous speech.

5. ANALISYS OF REAL SENTENCES

To find the above-mentioned relation we have used a subset of the Spanish database “ALBAYZIN” [4], called “minigeo”, containing 600 continuously spoken sentences, pronounced by six female and six male speakers and a vocabulary size of 200 words.

The sentences in this database have been automatically segmented (using a HMM-based speech recognition system). Then three parameters were estimated for all the sentences in the database: energy, fundamental frequency, and vowel duration.

We first estimated the energy curve of the sentences using a frame-by-frame schedule. Frames were overlapped in such a way that the frame length was 52 ms and the frame shift was 10 ms.

The estimation of the pitch curve (or, equivalently, the fundamental frequency) followed the same frame schedule. Using a Cepstral peaks detector, including a median filter to smooth the curve, we made the estimation of the pitch. Also, double values of pitch were detected and eliminated.

For duration, the vowels were considered, distinguishing also the sillables formed by diphtongs, and using the automatic segmentation as well as the transcription of the sentences.

Transcriptions were also used to ignore monosillabic words, leaving 2929 words to analyse. Rejecting the wrongly recognized words, the number of useful words for this study is 2851. Besides, theoretical accentual structure for the sentences was determined, according to orthographic rules and to considerations about grammatical function of words, as mentioned before.

Table I shows the histogram of different positions of the orthographic accent for the sentences in the database. In this figure H stands for a stressed syllable and L for a non-stressed syllable.

TABLE I
NUMBER OF DIFFERENT POSITIONS IN THE ORTHOGRAPHIC ACCENT.

Beginning with		Ending with	
H	1480	H	604
LH	671	HL	1979
LLH	383	HLL	266
LLLH	317	HLLL	2

Now we analyse the way the simple rules for stressed vowels in isolated words behave for continuous speech. As mentioned before, for isolated words, the values of energy, fundamental frequency and duration of stressed vowels is higher that the corresponding values for atonal vowels. This is estimated by counting the number of times the stressed vowel coincides with the maximum of those parameters along the word. This is shown in Table II.

It has some interest repeating this table with minima of fundamental frequency, instead of maxima. The results are shown in Table III, where it can be seen that there is a better agreement than in Table II.

TABLE II
MATCHING FOR LOCAL MAXIMA AND ORTHOGRAPHIC ACCENT.
E=ENERGY; F_0=FUNDAMENTAL FREQUENCY; D=DURATION.
✓=MATCH BETWEEN PARAMETER AND ACCENTUATION;
×=NO MATCH BETWEEN PARAMETER AND ACCENTUATION.

Maxima			% success
E	F_0	**D**	
×	×	×	17.71
×	×	✓	18.03
×	✓	×	4.60
×	✓	✓	8.19
✓	×	×	13.14
✓	×	✓	17.26
✓	✓	×	6.34
✓	✓	✓	14.68

TABLE III
MATCHING OF ENERGY (E) AND DURATION (D) MAXIMA BUT FUNDAMENTAL FREQUENCY (F_0) MINIMA AND ACCENTUATION.
✓=MATCH BETWEEN PARAMETER AND ACCENTUATION;
×=NO MATCH BETWEEN PARAMETER AND ACCENTUATION.

Max	min	Max	% success
E	F_0	**D**	
×	×	×	11.61
×	×	✓	11.82
×	✓	×	10.71
×	✓	✓	14.40
✓	×	×	12.07
✓	×	✓	16.56
✓	✓	×	7.43
✓	✓	✓	15.38

We found that removing those words immediately after or before a pause the agreement is improved in about 10%.

This and other series of tests allow us to conclude that although the relatively simple relation between prosodic features and stress we can find in isolated words is lost in continuous speech, there still exist some more complex relations to be discovered and used.

6. ESTIMATION OF THE ACCENTUAL STRUCTURE

The first process to be performed in a prosody-assisted recognition system is the estimation of the accentual structure of the sentence to be recognized, from the signal. From now on we suppress the duration as a parameter related to the stress. The reason is that we are interested in getting a one-pass system at the end. The use of duration would impose us the need of a segmentation pass, before the recognition itself. We made two classes of estimators: one base on HMM and other based on neural tree networks, NTN [13].

Estimation with HMM

Two different HMM acoustic models were trained (L and H). A recognition system was built using a bi-grammar of possible accentual structures. The parameterization was the same parameterization of the final ASR system, in which this information has to be embedded. After tuning the system, the rate of correct estimation of the prosodic structure (compared to the theoretical one) was 56.94 % in terms of words.

Estimation with NTN

Neural tree networks is an approach that combines the structure of decision trees with nodes that are pattern classifiers based on neural networks (Self-Organizing Maps, in this work). This technique lead to a rate of correct estimation of 85.65 % using only energy and pitch as parameters. The drawback of this method is that it requires the a priori segmentation of the sentence in terms of syllables.

According to the previous results, we present to kind of tests in the following sections. Tests based on the HMM estimator are taken as the worst case, in the sense that we hope that further improvements in the estimation method would result in a better rate of correct estimation of the accentual structures. This not seems to be difficult as the NTN estimator gets a clearly higher rate. On the other hand, tests based in the theoretical accentual structure are considered as the optimal estimation, providing a ceiling of what we should expect from our proposal.

7. THE GRAMMAR WITH PROSODY

One way of introducing prosodic information into the recognition process is to modify the grammar in such a way that the accentual structure of the sentence to be recognized is used to penalize those hypothesis that do not have the same structure.

For the sake of simplicity in the series of tests we have adopted a strategy consisting in three stages: estimating the accentual structure of the sentence, modifying the grammar to be used, and carrying out the recognition process itself. This strategy implies that the signal is used twice, eliminating the possibility of frame-synchronous recognition. But once the suitability of the approach is proved, we are now working in performing the whole process in only one pass.

Therefore, we now explain the way the grammar is modified. The recognition step is a regular one.

8. GRAMMAR PENALIZATION

The penalization step consists on modifying the probabilities of the bi-grammar according to the following penalization function, which multiplies the normal probability and takes the value:

1) Ke, when the hypothesis under evaluation has more words than the estimated structure.
2) 1+Ks D(w,a), where D(w,a) is a distance function between the hypothesized word, w, and its estimated prosodic structure, a. At the moment, the distance D is set to 0 if the estimated structure matches the word and 1 otherwise. This value of the penalization function is used when the word is the first or the last of the sentence.
3) 1+Kn D(w,a), for those transitions in the grammar whose probability was determined by a smoothing technique [15].
4) 1+Kw D(w,a), for the rest of transitions.

With the introduction of the penalization in the probabilities of the grammar, this probabilities not only depends on the history (previous words) but also on the accentual suitability.

9. IMPLEMENTATION

It is not adequate to use the bi-grammar in the usual way, as one particular transition form one word to another has to be penalized according to accentual structure, which is changes with the position of the words along the sentence.
To avoid this problem we build a new non-recursive finite state grammar that represents all possible transition of the bi-grammar for the first N words in

the sentence. It is like cutting all backward transitions and repeating the bi-grammar for N times. Provided the length of the sentence is not greater than N, this modified grammar behaves exactly like the original one. But now it can be penalized according to the position of the word within the sentence.

10. EXPERIMENTAL RESULTS

After tuning all penalization parameters, we obtain the following experimental result for a different partition of the database. We perform the recognition process for this partition in three different situations (Table IV): The first one, using the normal recognition system, as a reference. The second one includes prosodic information with an estimation of the accentual structure based on HMM. And finally the third one, which uses the theoretical accentual structure. We compute the Word Error Rate (WER) and the percentage of reduction of this WER for all cases. As it can be seen in Table IV, with the very realistic HMM estimation we obtain 28.91 % of reduction of the word error rate. The maximum improvement we should expect with this procedure (and the set of parameters we are using) is 36.87 %, obtained with the theoretical accentual structure.

TABLE IV

	WER %	%WER Reduction
Reference	7.54	-
HMM	5.36	28.91
Theoretical	4.76	36.87

11. REFERENCES

[1] Alarcos Llorach E., *Gramática de la Lengua Española*, Madrid: Espasa Calpe, pp. 52-68, 1999.

[2] Bartkova K. and Jouvet D., *Selective prosodic post-processing for improving recognition of French telephone numbers*, Proc. of 7th European Conference on Speech Communication and Technology, Vol. 1, pp. 267-270, 1999.

[3] Busdhtein D., *Robust Parametric Modeling of Durations in Hidden Markov Models*, IEEE Trans. On Speech and Audio Processing, Vol. 4, No. 3, 1996.

[4] Casacuberta F., García R., Llisterri J. Nadeu C., Prado J. M. and Rubio A., *Development of a Spanish Corpora for the Speech Research*, Workshop on International Co-operation and Standardisation of Speech Databases and Speech I/O Assessment Methods, CEC DGXIII, ESCA and ESPRIT PROJECT 2589 "SAM", Chiavari, 26-28 September 1991.

[5] Caspers J., *Testing The Meaning of Four Dutch Pitch Accent Types*, Proc. of 5th European Conference on Speech Communication and Technology, Vol. 2, pp. 863-866, 1997.

[6] Chiang T-H, Lin Y-C and Su K-Y, *On Joinly Learning the Parameters in a Character Synchronous Integrated Speech and Language Model*, IEEE Trans. On Speech and Audio Processing, Vol. 4, No. 3, 1996.

[7] Chih-Heng L., Chien-Hsing W., Pei-Yih T. and Hsin-Min W., *Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units*, Speech Communication, Vol. 18, pp. 175-190, 1996.

[8] Deller, J. R., Proakis J. G., Hansen J. H., *Discrete-Time Processing of Speech Signals*, Prentice Hall, 1987.

[9] Hirose K. and Iwano K., *Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition*, Proc. of IEEE 25rd International Conference on Acoustics, Speech and Signal Processing, Vol. 3, pp. 1763-1766, 2000.

[10] Lee S-W. and Hirose K. *Dynamic beam-search strategy using prosodic-syntactic information*, Workshop on Automatic Speech Recognition and Understanding, pp. 189-192, 1999.

[11] Lee T. and Ching C., *Cantonese Syllable Recognition Using Neural Networks*, IEEE Trans. On Speech and Audio Processing, Vol. 7, No. 4, 1999.

[12] López E., Caminero J., Cortázar I. and Hernández L., *Improvement on Connected Numbers Recognition Using Prosodic Information*, Proc. of 5th International Conference on Spoken Language Processing, Prosody and Emotion 2, 1998.

[13] Milone D. H., Sáez J. C., Simón G., Rufiner H. L., *Self-Organizing Neural Tree Networks*, Proc. of 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 20, No. 3, 1998.

[14] Molloy L. and Isard S., *Suprasegmental Duration Modeling with Elastic Contraints in Automatic Speech Recognition*, Proc. of 5th International Conference on Spoken Language Processing, Hidden Markov Model Techniques 3, 1998.

[15] Potamianos G. and Jelinek F., *A study of n-gram and decision tree letter language*

modeling methods, Speech Communication, Vol. 24, pp. 171-192, 1998.

[16] Quilis A., *Tratado de fonología y fonética españolas*, Madrid: Editorial Gredos, 1993.

[17] Rossi M., *Is Syntactic Structure Prosodically Retrievable?,* Proc. of 5th European Conference on Speech Communication and Technology, Keynote Speech, 1997.

[18] Sönmez M. K., Heck L., Weintraub M., Shriberg E., *A Lognormal Tied Mixture Model of Pitch for Prosody Based Speaker Recognition*, Proc. of 5th European Conference on Speech Communication and Technology, Vol. 3, pp. 1391-1394, 1997.

[19] Stolcke A., Shriberg E., Hakkani-Tür D. and Tür G., *Modeling the prosody of hidden events for improved word recognition*, Proc. of 7th European Conference on Speech Communication and Technology, Vol. 1, pp. 311-314, 1999.

[20] Vereecken H., Vorstermans A., Martens J. P. and Van Coile B., *Improving the Phonetic Annotation by Means of Prosodic Phrasing*, Proc. of 5th European Conference on Speech Communication and Technology, Vol. 1, pp. 179-182, 1997.

[21] Wang C. and Seneff S., *A Study of Tones and Tempo in Continuous Mandarin Digit Strings and Their Application in Telephone Quality Speech Recognition*, Proc. of 5th International Conference on Spoken Language Processing, Prosody and Emotion 2, 1998.

[22] Warnke V., Gallwitz F., Batliner A., Buckow J., Huber R., Nöth E. and Höthker A., *Integrating Multiple Knowledge Sources for Word Hypotheses Graph Interpretation*, Proc. of 7th European Conference on Speech Communication and Technology, Vol. 1, pp. 235-238, 1999.

[23] Wu S-L., Kingsbury B. E. D., Morgan N., Greenberg S., *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*, Proc. of IEEE 23rd International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 721-724, 1998.

[24] Yaeger-Dror M., *Register as a variable in prosodic analysis: The case of the English negative*, Speech Communication, Vol. 19, pp. 39-60, 1996.