N-BEST RESCORING BASED ON INTONATION PREDICTION FOR A SPANISH ASR SYSTEM

Diego A. Evin, Jorge A. Gurlekian, Humberto M. Torres

Laboratorio de Investigaciones Sensoriales, Hospital de Clínicas, University of Buenos Aires

diegovin@gmail.com, jag@fmed.uba.ar, hmtorres@hotmail.com

Abstract: This paper presents a novel method for rescoring the n-best recognition hypotheses using intonation knowledge. The model synthesizes the f0 contours for each of the n-best hypotheses and estimates an intonative matching index between the synthetic shapes and the real f0 contour. This index is applied in the rescoring process, and can be viewed as a degree of intonation compatibility between the hypotheses and the input sentence. The f0 prediction is based on classification and regression trees and the Fujisaki model. We evaluate our approach using a single speaker of the Buenos Aires Spanish LIS-SECYT database under clean and babble-noisy conditions. Considering the systems under no grammar condition, the proposed model reduces the mean absolute word error rate in 3.1% with respect to the baseline system, in a consistent manner and under different noise conditions.

1 Introduction

In the searching for automatic speech recognition (ASR) systems able to reach the human recognition capabilities, it is often analyzed which knowledge sources and processing strategies used by humans are still underexploited in the standard ASR framework.
Between the additional knowledge sources that the listeners are able to integrate during the process of decoding an oral message, have been reported: syntactic, semantic, pragmatic and prosodic information. Prosodic information in particular, is related with the others linguistic levels of speech processing.

A number of papers have analyzed and experimented with the incorporation of prosodic information into ASR systems. We can find, for example, proposals about the use of prosodic information at the acoustic phonetic level, in the discrimination of voiced and unvoiced stops; at the lexical level to detect the accent position in the word; at syntactic level to locate the principal phonological boundaries in sentences; and at the pragmatic level to locate the emphatic portions of discourse, since more than three decades ago [1].

We can also find experiments applying prosodic information in every specific stage of the standard ASR process. It is, during the preprocessing (for example segmenting the incoming speech into sentences and phrases); as an additional acoustic cue in the observation features vector (during the training of acoustic models); as a supplementary evidence during the search process; and in the post-processing stage, in the reestimation of recognition hypotheses [2].

To cite just a few antecedents in using prosodic information for n-best rescoring we can mention the work of [3], where two different alternatives are proposed. In first term they propose to use a word duration model, based on the individual duration of constituent phones. In second place they study the use of a predictive pause model from the inter-word context and the application of that information jointly with the patterns of pauses of the input signal. In this work a reduction in absolute word error rates (WER) from 0.2% to 0.9% are reported.

In [4] it is presented a method for the prediction of word level durations using a probability density function, and the integration of this model to rescoring the n-best hypotheses. Test presented shows a WER reduction of 0.2% for English and 0.1% for Spanish.

In [5] the authors propose a system based on HMM for the prosodic segmentation of the input speech into phonological phrases for fixed stress languages. Word boundaries information is derived from the prosodic boundaries, and is integrated into the rescoring process of an ASR system. Using this rescoring strategy, the authors obtained a relative improvement of the ratio of correctly recognized words by 3.82% for a Hungarian medical speech recognition task.

Making use of the observation that for a given speaker, sentence, and speech rate, a certain type of intonation pattern is more probable to find than others, we studied the feasibility of using a measure of intonative compatibility between the recognition hypotheses and the pattern observed in the acoustic input signal, with the aim to improve the performance of a standard ASR system.

The following part of the paper is constructed as follows; in section 2, the proposed scheme of N-best hypotheses rescoring is explained. Section 2.1 outlines the proposed model, section 2.2 describes the intonation prediction module as well as the procedure used for input features extraction. Evaluation experiments and results are presented in section 3. Finally section 4 exposes the conclusions and the implications of the results for future works.

2 Intonation Based N-Best Rescoring

2.1 Outlines

As was stated previously, the proposed system relies on the assumption that for a given speaker, sentence and speech rate, certain types of prosodic patterns are more likely to occur than others. The system tries to exploit that correlation between sentences and intonation as additional evidence during the automatic speech recognition. A summary of the proposed strategy is presented in Figure 1.

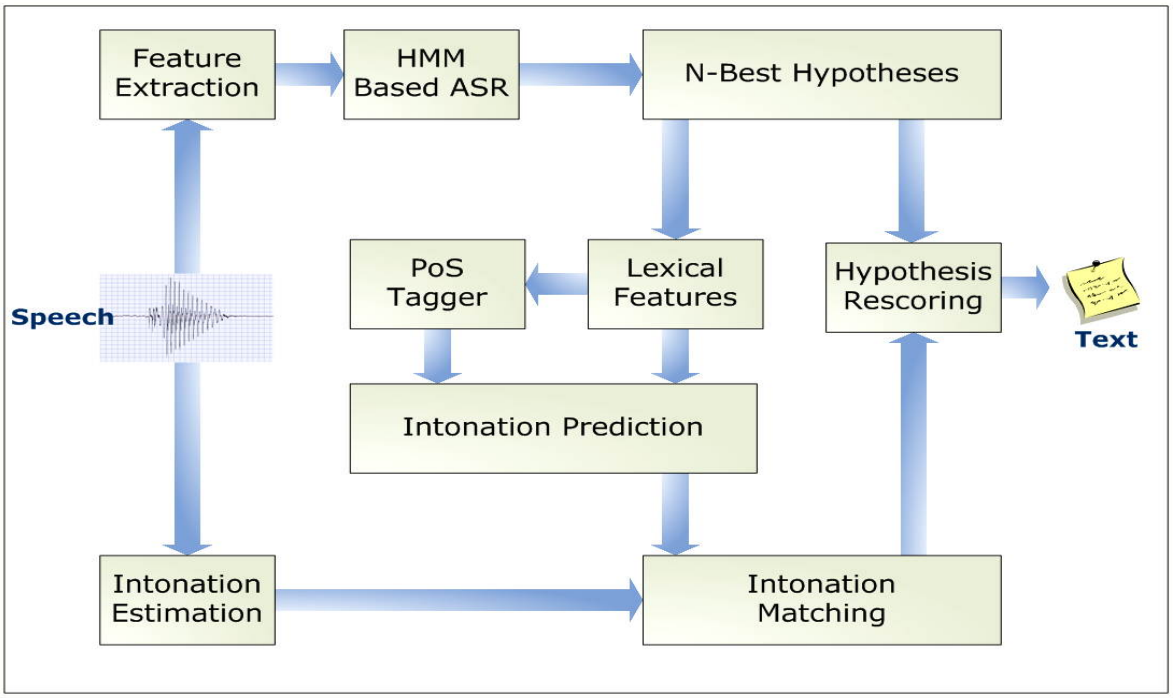

Figure 1 - General scheme of the proposed ASR system

The signal to be recognized is received by the feature extraction module, which in turn derives the set of speech parameters to a conventional HMM ASR system. This ASR system outputs a list of N-Best recognition hypotheses. Once the set of hypothesized sentences is available, a lexical analysis is performed. Using forced alignment, the phone and word level segmentation is obtained, and through a part-of-speech (POS) tagging [6] the POS labels are extracted for each candidate sentence. The lexical features module extracts and provides to the intonation prediction module the required information for the generation of a single f0 contour per hypothesis. This module will be explained more deeply in the next section

The input speech signal is also analyzed to extract the f0 contour (target intonation curve), using the algorithm getF0 [7], in the intonation estimation module.

Finally, the intonation matching module compares the target f0 contour against each of the candidates and generates an intonation compatibility index. There are several algorithms proposed to compare two time series. In this work we used the fact that the target f0 as well as each of the candidates has the same length. Furthermore, as we are interested in comparing in a point to point basis each pair of f0 curve, time warping and other adaptation techniques for comparison are not essential. We evaluated three different measures to test the difference between the corresponding contours: mean square error, overall error in Hz, and overall error in ERBs.

The process of hypothesis rescoring consists in the combination of the original likelihood and the intonation compatibility index to get a new ranking of recognition hypotheses.

2.2 Intonation Prediction

This module resembles the prosodic module of standard concatenative text to speech systems (TTS). Two sub-modules can be distinguished here: the Fujisaki commands predictor and the f0 synthesizer.

We decided to use an intermediate representation to simplify the process of f0 curve prediction. Using a parametric representation, the f0 prediction problem consists in finding the values of such parameters instead of the whole waveform. We choose to use the Fujisaki superpositional model of f0 [8] that was already tested [9] for Argentine-Spanish and compared [9] with a phonological approach using linguistics marks as in ToBI. This model analytically describes the f0 contour in a log scale, as the superposition of three components: a base frequency, the tonal accents and the phrase accents. Phrase accents are calculated as the response to a second order lineal filter critically damped, excited with a delta function called phrase command. Tonal accents resulted from the response to the same filter, excited with a step function called accent command.

$$\ln f0(t) = \ln(f_{\min}) + \sum_{i=1}^{N_f} Af_i Gf_i(t - T0_i) + \sum_{j=1}^{N_a} Aa_j \left\{ Ga_j(t - T1_j) - Ga_j(t - T2_j) \right\} \qquad \textbf{(1)}$$

where $Gf_i(t)$ is given as:

$$Gf_i(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t} & t \geq 0 \\ 0 & t < 0 \end{cases} \qquad \textbf{(2)}$$

and $Ga_j(t)$ is given as:

$$Ga_j(t) = \begin{cases} \min\left\{ 1 - (1 + \beta_j t) \cdot e^{-\beta_j t}, \quad \gamma_j \right\} & t \geq 0 \\ 0 & t < 0 \end{cases} \qquad \textbf{(3)}$$

Parameters α and β in equations 2 and 3 characterize the dynamic properties of the laryngeal mechanisms of phrase and accent control. Together with γ they can be considered practically constant for all speakers. The variable f_{min} in equation 1 must be estimated for each emission. Finally, the parameters to be calculated are the phrase commands, amplitude values of the phrase accents Af, amplitude values of tonal accents Aa and times T0, T1 and T2.
Because of the complex formulation of the model, analysis and automatic parameter extraction from f0 measurements is time consuming and needs several manual revisions. One of the standard methods is analysis-by-synthesis [11]. This requires a complete search by quantified steps, within a reasonable range for each parameter. The iterative process continues until the best value combination fits the measured contour.

Fujisaki commands were predicted using classification and regression trees (CARTS) using the algorithm proposed by Breiman et al [12].
At the training phase, different input types were used to feed the trees:

- Location and length of the intonative phrase.
- Identity, location and length of the stressed vowel.
- Identity of the context phonemes, two previous and two posterior phonemes.
- Part of Speech (POS), in a context of two words before and two after.
- Distance to the previous and following stressed vowel that has an accent command associated.
- Parameter values of the previous accent and/or phrase command.

For inputs corresponding to location, distance and length, different measurement units were used: time and number of phonemes, syllables, words and intonative phrases. These measurements were introduced as absolute and relative values.

Parameters to be estimated are listed below:

- **Af:** phrase command amplitude.
- **$T0_r$:** phrase command location, relative to the beginning of the intonative phrase.
- **Aa:** accent command amplitude.
- **$T1_r$:** accent command location relative to stressed vowels of content words, except the last one which does not have any associated command.
- **T2-T1:** accent command duration.

Taking into account the high number of degrees of freedom in the combinations of Fujisaki commands that can fit an f0 contour, in the construction of the trees we used the following assumptions: we admit just one phrase command per sentence; furthermore we assume one accent command for each content word, except the last content word in the sentence.
We trained a tree for each Fujisaki parameter. To evaluate the performance of each predictor, cross validation was employed, dividing randomly the 80% of the data for training (used to estimate the CARTs), and the other 20% for testing. The percentage of data designated for training is a critical factor, because CART performance is very sensitive to the quantity and quality of the data used to estimate the model. If the percentage to test is too low, a poor generalization will be obtained, and if it is too high there will not be enough data to train the models [13].
The experiments were done over five partitions of the available data set and the results were averaged. For different data sets, a series of experiments were performed to find CART's parameters that minimize the classification error. The best balance was found for a tree depth of three. Table 1 shows the mean percentage RMSE error during the estimation of the Fujisaki

model parameters over all the partitions using the optimum tree depth. The RMSE errors presented are relative to the values of the Fujisaki parameters estimated using genetic algorithms, as described in [14].

Parameter	Af	Aa	T0r	T1r	T2 -T1
RMSE	8%	18%	31%	25%	29%

Table 1 - Mean RMSE percentages predicting the Fujisaki model parameter

During the CART's construction, the method automatically selected the inputs to be used. The features that were chosen as inputs for each of the predicted parameters are listed below:

- **Af:** identity of the previous and next phonemes, distance from the center of the accented vowels of the firs content word to the end of the phrase in seconds, duration of the previous accent command in seconds.
- **$T0_r$:** stressed vowel location of the first content word in seconds, phrase length in seconds, duration of the previous accent command in seconds, previous phrase command amplitude.
- **Aa:** identity of the previous and next phonemes, next phoneme articulation mode, part of speech (POS) label for the current, previous, next and after next content words, number of intonative phrases in the sentence, phrase length in seconds.
- **$T1_r$:** identity of the previous and next phonemes, distance from the stressed vowel of the current content word to the end of the intonation phrase in seconds.
- **T2-T1:** identity of the next phoneme, phrase length in seconds, distance from the stressed vowel of the current content word to the beginning of intonation phrase in seconds, distance from the stressed vowel of the current content word to the end of the intonation phrase in seconds, duration of the previous accent command in seconds.

It is worth mentioning that for the phrase command parameters, each one of the detailed input features are relative to the location of the stressed vowel of the first content word. While for the accent commands, the measures of the features are successively calculated with respect to the location of the stressed vowel of the current content word. As was previously mentioned, a command accent is associated with each content word (except the last one).

An information that results interesting to know, to use the module for intonation prediction into the proposed rescoring strategy, is how well the predicted f0 fits to the observed f0 values. In Table 2, the results of objective evaluations, using different measures of errors, between the predictions of f0 using the Fujisaki model and CART, and the actual values are presented. The different frequency scales of the presented figures are used in the next section to compare the matching of the recognition hypothesis.

RMS (Hz)	RMSE (ST)	RMSE (Ln)	RMSE (ERBs)	MSE (Ln)
64	4.7	0.27	1.12	0.078

Table 2 - Mean errors with the f0 prediction model. ST: Semitones; Ln: logarıthmic scale

Where RMSE is measured as the difference between the original and the predicted f0, considering only voicing segments without interpolation of unvoiced segments. MSE values were also calculated with f0 in the logarithmic domain to make it comparable with [15].

Once the Fujisaki parameters are predicted, the f0 synthesizer makes use of equations 1-3 to generate the f0 contour.

3 Evaluation Experiments

3.1 Corpus

The data used as case of study for the proposed model was obtained from the LIS-SECYT [16] database. It consists of 741 declarative sentences extracted from Buenos Aires newspapers. The sentences have from 1 to 5 intonative phrases and contain 97% of all Spanish syllables, in both stress conditions and all possible syllable positions within the word. Two native speakers female and male, read the sentences in a sound proof chamber. Recordings were made with an AKG dynamic microphone and 16 kHz/16bit conversion. The speakers were instructed to read the sentences with all kind of natural tonal variations.

Each sound file was manually labeled twice, by musically trained speech therapists who distinguished prosodic occurrences as intonational groups and tonal accents. The files were labeled in different layers: phonetic, orthographic, breaks and tonal marks according to a phonetic method [9] Parts of speech and syntactic layers were also indicated.

Analysis of phonological labels for [17] revealed that 84% of content words receive a tonal accent of type H* and 10% receive a delayed peak as L*+H. That is Arg-Spanish has a high tonal accent in almost every content word in read speech but not in the last one of the final phrase.

3.2 Reference ASR system

In order to evaluate the proposed strategy we constructed a reference ASR system (baseline system). In addition to this baseline recognizer, the system was extended by the intonation based N-best rescoring module outlined in 2.1.

Under the hypothesis of higher robustness of the suprasegmental information over spectral information, the proposed system was evaluated on clean and noisy conditions using the LIS-SECYT database. The noisy version was created by adding babble noise at 5, 10, 15 and 20dB SNR (signal-to-noise ratio) to the original clean utterances. Babble noise was obtained from the NOISEX database [18]. For each SNR, model sets were trained and evaluated under matched (training/testing) conditions. It means that to evaluate for example the test set of 5dB SNR, the corresponding training data of 5dB SNR were used to construct the models. The experiments were developed under 10-fold cross validation methodology, using 80% of the data for training and 20% for testing.

The Hidden Markov Model Toolkit (HTK) [19] was used for the speech recognition experiments. Each model was represented by a continuous density HMM with left-to-right configuration. The speech data was pre-emphasized, windowed to 25-ms frames with 10-ms frame shift, and parameterized into 39 order MFCCs, consisting of 12 cepstral coefficients, energy, and their first and second order differences. Since the size of the dataset available, 32 monophones were employed as acoustic models. In all the evaluations we have used no grammar, to better observe the improvements due to the proposed strategy and no to the language model.

Once the list of N-best hypothesis for each sentence is obtained, as was explained, the proposed model requires determining the set of input features for the f0 prediction module. These sets of features are determined as described in section 2.1. Furthermore, an f0 prediction model was trained for the whole dataset.

After that, a predicted f0 curve is obtained for each hypothesis using the model explained in section 2.2. Using the intonation matching between the actual f0 curve and the predicted ones, the new N-best hypothesis is shaped. In the entire test sets, the best intonation matching (i.e. the measure that shows the minimum overall word error rates) was the RMSE in Hz, and was the frequency scale used in the reported results.

3.3 Results

Table 3, shows the performance comparison between the baseline and the proposed systems in terms of percentage of mean word error rate (WER) for the 10 corresponding folds, under different SNRs. In all the cases the proposed model was applied using 5-best rescoring.

Model	Condition	%WER
Baseline	Clean	23.43
Proposed System		19.56
Baseline	Babble 20 dB	25.27
Proposed System		22.66
Baseline	Babble 15 dB	30.41
Proposed System		27.24
Baseline	Babble 10dB	39.12
Proposed System		35.77
Baseline	Babble 5 dB	54.94
Proposed System		52.47

Table 3 - Recognition performance comparison between the baseline and proposed systems

It can be seen that the proposed system improves the performance of the baseline, no matter the SNR level of the audio. The proposed system reduces the absolute percentage of WER in 3.094% mean approximately with respect to the baseline system.

4 Conclusions

This work shows a novel methodology to make use of intonation knowledge in the rescoring of recognition hypothesis. The proposed method showed to be useful improving the performance of a baseline ASR system for Spanish, speaker dependent task under no grammar condition. However it must be noted that the proposed model used a prediction module trained on the same dataset, which can imply the requirement of fitting that module for each new speaker to reach this results. It can seem logic to think that the shape of intonation contour for a given sentence is speaker dependant.

It must be noted that the same technique can also be extended to use other prosodic features during the disambiguation process.

As future works, it remains to evaluate the computational overhead that implies the proposed method over the baseline; the result obtained using language models, and the behavior under

multi speaker scenario. Also we are planning to extend this idea using an intonation prediction module that outputs a set of candidate intonation contours instead of just one of them.

5 Acknowledgements

The authors would like to thank to CONICET and MINCYT for their financial support.

References

[1] Lea, W.: Prosodic Aids to Speech Recognition, pp. 66-205 in (W. Lea, ed.) Trends in Speech Recognition, Prentice-Hall, Englewood Cliffs, New Jersey, 1980.
[2] Cutler, A. : Dahan, D., van Donselaar, W., Prosody in the Comprehension of Spoken Language: A Literature Review. Language and Speech, 40 (2), pp. 124-201, 1997.
[3] Vergyri, D. : Stolcke, A., Gadde, V., Ferrer, L., and Shriberg, E., Prosodic Knowledge Sources for Automatic Speech Recognition. In Proc. ICASSP, vol. 1, pp. 208–211, Hong Kong, 2003.
[4] Seppi, D.: Falavigna, D., Stemmer, G., and Gretter, R., Word Duration Modeling for Word Graph Rescoring in LVCSR. In Proc. INTERSPEECH, pp. 1805–1808. Antwerp, Belgium 2007.
[5] Vicsi, K. : and Szaszak, G., Using Prosody to Improve Automatic Speech Recognition. Speech Communication (52) 413–426. Elsevier 2010.
[6] Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., and Padró, M.: FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library. In Proc. of the 5th International Conference on Language Resources and Evaluation (LREC 2006), ELRA. Genoa, 2006.
[7] Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). In W. B. Klejin, & Paliwal, K. K (Ed.), Speech Coding and Synthesis, pp. 495-518. Elsevier, Amsterdam 1995.
[8] Fujisaki, H., and Hirose, K.: Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese. Journal of Acoustic Society 5 (4), 233–242, 1984.
[9] Fujisaki, H., Ohono, S., Ichi Nakamura, K., Guirao, M., and Gurlekian, J.: Analysis of Accent and Intonation in Spanish Based on a Quantitative Model. In: Proc. of ICSLP 94, pp. 355–358. Yokohama 1994.
[10] Gurlekian, J., Torres, H., and Colantoni, L.: Evaluación de las Descripciones Analítica y Perceptual de la Entonación de una Base de Datos de Oraciones Declarativas de Foco Amplio para el Español Hablado en Buenos Aires. Estudios de Fonética Experimental XIII: 275-302, 2003.
[11] Mixdorff, H.: A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. In: Proc. of ICASSP 2000. Vol. 3, pp. 1281–1284. Istanbul, 2000.
[12] Breiman, L.: Friedman, J., and Stone, C., Clasification and Regression Trees. Chapman & Hall, New York, 1984.
[13] Goutte, C.: and Larsen, J., Optimal Cross - Validation Split Ratio: Experimental Investigation. In Proc. of ICANN'98, pp. 681–686. Skövde, Sweden, 1998.
[14] Torres H. and Gurlekian J.: Parameter Estimation and Prediction from Text for a Superpositional Intonation Model. In Proc. of the 20 Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2009), pp. 238-247, TUDpress Verlag der Wissenschaften, Dresden, Germany, September 21-23, 2009.
[15] Sakurai, A., K, H., and Minematsu, N.: Data-driven generation of F0 Contours Using a Superpositonal Model. Speech Communication 40 (4), 535–549. Elsevier 2003.
[16] Gurlekian, J. A., Rodríguez, H., Colantoni, L., and Torres, H. M.: Development of a Prosodic Database for an Argentine Spanish Text to Speech System. In Proc. of IRCS Workshop on Linguistic Databases, Philadelphia, 2001.
[17] Colantoni, L., and Gurlekian, J.: Convergence and Intonation: Historical Evidence from

Buenos Aires Spanish. Bilingualism: Language and Cognition. Penn State Univ. Vol.7, Nr.2, pp. 107-119, Aug. 2004.
[18] Varga, A., and Steeneken, H.: Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. Speech Communication, 12(3):247-251, Elsevier 1993.
[19] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P.: The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department, 2006.