Prosodically-based automatic segmentation and punctuation

Helena Moniz[1,2]*, Fernando Batista*[2,3]*, Hugo Meinedo*[2]*, Alberto Abad*[2]*,*
Isabel Trancoso[2]*, Ana Isabel Mata*[1]*, Nuno Mamede*[2]

[1]FLUL/CLUL, University of Lisbon, Portugal
[2]IST / INESC-ID, Lisbon, Portugal
[3]ISCTE, Lisbon, Portugal

{helenam;fmmb;meinedo;alberto;isabel.trancoso;njm}@l2f.inesc-id.pt and aim@fl.ul.pt

Abstract

This work explores prosodic/acoustic cues for improving a baseline phone segmentation module. The baseline version is provided by a large vocabulary continuous speech recognition system. An analysis of the baseline results revealed problems in word boundary detection, that we tried to solve by using post-processing rules based on prosodic features (pitch, energy and duration). These rules achieved better results in terms of inter-word pause detection, durations of silent pauses previously detected, and also durations of phones at initial and final sentence-like unit level. These improvements may be relevant not only for retraining acoustic models, but also for the automatic punctuation task. These two tasks were evaluated. Results based on more reliable boundaries are promising. This work allows us to tackle more challenging problems, combining prosodic and lexical features for the identification of sentence-like units.
Index Terms: prosody, automatic phone segmentation, punctuation.

1. Introduction

The main motivation of our work is the improvement of the punctuation module of our automatic broadcast news captioning system. Although this system is deployed in several languages, including English, Spanish and Brazilian Portuguese, the current paper refers only to the European Portuguese (EP) version. Like the audio diarization and speech recognition modules that precede them, the punctuation and capitalization modules share low latency requirements.

Although the use of prosodic features in automatic punctuation methods is well studied for some languages, the first implemented version for EP deals only with full stop and comma recovery, and explores a limited set of features, simultaneously targeting at low latency, and language independence. The aim of this work is to improve the punctuation module, first by exploring additional features, namely prosodic ones, and later by encompassing interrogatives. This paper describes our steps in this first direction.

One of the most important prosodic features is the duration of silent pauses. Even though they may not be directly converted into punctuation, silent pauses are in fact a basic cue for punctuation and speaker diarization. The durations of phones and silent pauses are automatically provided by our large vocabulary continuous speech recognition module. An analysis of these results, however, revealed several problems, namely in the boundaries of silent pauses, and in their frequent miss-detection. These problems motivated the use of post-processing rules based on prosodic features, to better adjust the boundaries of silent pauses. The better results achieved with this prosodic module motivated the retraining of both the acoustic models and punctuation models.

This work was done using a subset of the EP broadcast news corpus, collected during the ALERT European project. Although the corpus used for training/development/evaluation of the speech recognizer includes 51h/5h of orthographically transcribed audio, a limited subset of 1h was transcribed at the word boundary level, in order to allow us to evaluate the efficacy of the post-processing rules. With this sample we could evaluate the speech segmentation robustness with several speakers in prepared non-scripted and spontaneous speech settings with different strategies regarding speech segmentation and speech rate.

The next section reviews related work. The post-processing rules and their impact on word segmentation results are described in Sections 3 and 4, respectively. Section 5 deals with the retraining of acoustic models. Section 6 is devoted to the description of the punctuation module and the results obtained before and after the retraining. Conclusions and future work are presented in Section 7.

2. Related work

2.1. Segmentation

Speech has flexible structures. Speaker plans what to say online and makes use of different cues from context, thus spontaneous speech may have elliptic utterances, backchannel expressions, disfluencies, and overlapping speech, *inter alia*. It is also characterized by temporal characteristics (speech rate, elongated linguistic material, etc) that make its modeling difficult. This set of properties poses interesting challenging both from a linguistic and from an automatic speech recognition point of view [1]. Our work focus in broadcast news, where one can find both spontaneous and read speech.

The study by [2] shows that there are different levels of segmentation when combining linguistic features with automatic methods. We could say that those different types of segmentation may reflect an increasing gradient scale: pause-based, pause-based with lexical information, the previous and dialog acts information, topic and speaker segmentation. While pause and speaker segmentation are based on audio diarization techniques, the remaining types are related with structural segmentation methods. Audio diarization may comprehend identification of jingles, speech/non-speech detection, speaker clustering and identification, etc. Structural segmentation concerns algorithms based on linguistic information to delimit "spoken sentences" (units that may not be isomorphic to written sentences),

and topic and story segmentation. This structural segmentation is the core of our present work.

Several studies (e.g., [3, 1, 4, 5, 2, 6]) have been showing that the analysis of prosodic features is used to model and improve natural language processing systems. The set of prosodic features, such as pause, final lengthening, pitch reset, *inter alia*, are among the most salient cues used in algorithms based on linguistic information. The implementation is supported on evidences that this cues are language-independent [7] and also on the fact that already studied languages have prosodic strategies to delimit sentence-like units (SU) and paragraphs with pitch amplitude, pitch contours, boundary tones and pauses.

We do know that there is no one-to-one mapping between prosody and punctuation. Silent pauses, for instance, can not be directly transformed in punctuation marks for different reasons, e.g. prosodic constraints regarding the weight of a constituent; speech rate; style; different pragmatic functions, such as emphasis, emotion, on-line planning. However, the correct identification of silent pauses and phone delimitation do contribute to the segmentation of speech in sentence-like units and do in fact contribute to punctuation.

2.2. Punctuation

Although different punctuation marks can be used in spoken texts, most of them rarely occur and are quite difficult to automatically insert or evaluate. Hence, most studies focus either on *full stop* or on *comma*. *Comma* is usually the most frequent punctuation mark, but it is also the most problematic because it serves many different purposes. [8] describes a method for inserting *commas* into text, and presents a qualitative evaluation based on the user satisfaction, concluding that the system performance is qualitatively higher than the sentence accuracy rate would indicate.

Detecting positions where a punctuation mark is missing, roughly corresponds to the task of detecting a SU, or finding the SU boundaries. SU boundary detection has gained increasing attention during recent years, and it has been part of the NIST rich transcription evaluations. A general HMM (Hidden Markov Model) framework that allows the combination of lexical and prosodic clues for recovering *full stop*, *comma* and *question marks* is used by [9] and [10]. A similar approach was also used for detecting sentence boundaries by [11, 1, 12]. [10] also combines 4-gram language models with a CART (Classification and Regression Tree) and concludes that prosodic information highly improve the results. [13] describes a maximum entropy (ME) based method for inserting punctuation marks into spontaneous conversational speech, where the punctuation task is considered as a tagging task and words are tagged with the appropriate punctuation. It covers three punctuation marks: *comma*, *full stop*, and *question mark*; and the best results on the ASR output are achieved using bigram-based features and combining lexical and prosodic features. [6] proposes a multipass linear fold algorithm for sentence boundary detection in spontaneous speech, which uses prosodic features, focusing on the relation between sentence boundaries and break indices and duration, covering their local and global structural properties. Other recent studies have shown that the best performance for the punctuation task is achieved when prosodic, morphologic and syntactic information are combined [2, 12, 14].

3. Word boundaries and silent pauses

3.1. Baseline phone segmentation

The first module of our broadcast news processing pipeline, after jingle detection, performs audio diarization [15]. The second module is the automatic speech recognition module. Audimus is a hybrid automatic speech recognizer [15] that combines the temporal modeling capabilities of Hidden Markov Models with the pattern discriminative classification capabilities of Multi-layer Perceptrons (MLP). The vocabulary has 100k words. Modeling context dependency is a particularly hard problem in hybrid systems. Our current solution uses, in addition to monophone units modeled by a single state, multiple-state monophone units, and a fixed set of phone transition units aimed at specifically modeling the most frequent intra-word phone transitions [16]. Using this strategy, the word error rate (WER) for the current test set of 5h was 22.0%.

This recognizer was used in a forced alignment mode in our reduced test set of 1h duration that was manually transcribed at the word boundary level. As explained above, this revealed several problems, namely in the boundaries of silent pauses, and in their frequent miss-detection.

3.2. Post-processing rules

Reducing these problems was the motivation for first applying post-processing rules to the baseline results, and later retraining the speech recognition models. These post-processing rules were applied off-line, and used both pitch and energy information. Pitch values were extracted using the Snack Sound Toolkit[1], but the only used information was the presence or absence of pitch.

The energy information was also extracted off-line for each audio file. Speech and non-speech portions of the audio data were automatically segmented at the frame-level with a bi-Gaussian model of the log energy distribution. That is, for each audio sample a 1-dimensional energy based Gaussian model of two mixtures is trained. In this case, the Gaussian mixture with the "lowest" mean is expected to correspond to the silence or background noise, and the one with the "highest" mean corresponds to speech. Then, frames of the audio file having a higher likelihood with the speech mixture are labeled as speech and those that are more likely generated by the non-speech mixture are labeled as silence.

The integration of extra info was implemented as a post-processing stage with three rules:

1. if the word starts by a plosive sound, the duration of the preceding pause is unchanged (typically around 50 to 60 ms for EP);
2. if the word starts or ends by a fricative, the energy-based segmentation is used;
3. if the word starts with a liquid sound, energy and pitch are used.
4. otherwise, they are delimited by pitch.

With these rules, we expected to have more adequate word boundaries than with our previous segmentation methods, without imposing thresholds for silent pause durations, recognized by [5] as misleading cues that do not account for differences between speakers, speech rate or speech genres.

[1] http://www.speech.kth.se/snack/

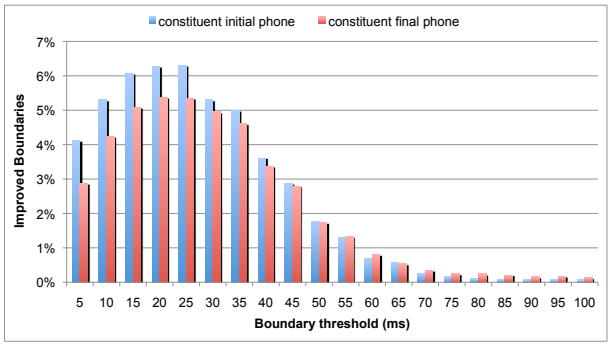

Figure 1: Improvement in terms of correct word boundaries, after post-processing.

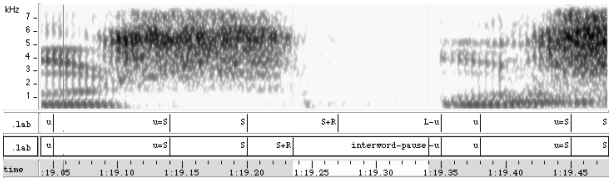

Figure 2: Phone segmentation before (top) and after (bottom) post-processing. The original sentence is "o Infarmed analisa cerca de quinhentos [medicamentos], os que levantam mais dúvidas quanto à sua eficácia." "Infarmed analysis about five hundred [drugs], the most doubtful about their effectiveness.". Initial and final word phones are marked with "L-", and "+R", respectively, whereas frequent phone transition units are marked with "=".

4. Results

By comparing the results in terms of word boundaries before and after the post-processing stage in the limited test set of 1h duration, we have found that 9.3% of the constituent initial phones and 10.1% of the constituent final phones were modified, in terms of boundaries. In what concerns the inter-word pauses, 62.5% of them were modified and 10.9% more were added. Figure 1 illustrates the improvement in terms of correct boundaries, when different boundary thresholds are used. The graph shows that most of the improvements are concentrated in an interval corresponding to 5-60 ms. Our manually reference has 443.82 seconds of inter-word pauses, the modified version correctly identified more 67.71 seconds of silence than in the original one, but there are still 14.81 seconds of silence that were not detected.

Figure 2 shows an example of a silent pause detection corresponding to a comma. The two automatic transcriptions correspond to the results obtained before (miss-detection) and after post-processing.

Two properties of EP trigger erroneous segmentation: phonetic fricatized voiced and unvoiced plosives [17], such as [d] or [t], and the epenthetic vowel (in EP [ɨ]), both at the end of a chunk followed by a pause. In EP, the last postonic syllable of a chunk corresponding to plosive and vowel may have the vowel elided and the remaining plosive uttered with fricatization effects - with a burst and unvoiced fricative spectra. To the best of our knowledge, the relationship of this effect with the prosodic structure is still not well known. In our reduced data set, these fricatization effects seem to consistently occur

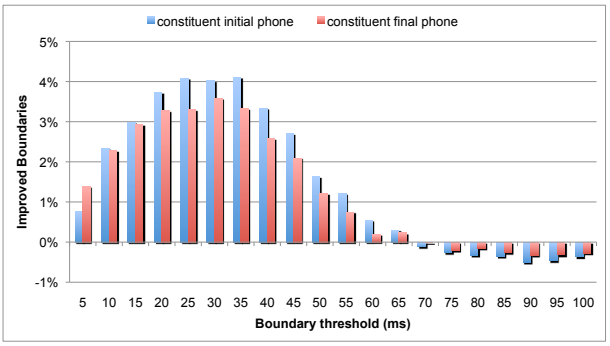

Figure 3: Improvement in terms of correct word boundaries, after retraining.

at the end of an intonational phrase and may phonetically indicate prosodic phrasing. The epenthetic vowel also causes miss-detections. The automatic segmentation method covers cases of reduction (one of the most frequent processes in EP), but it is still not well adjusted to this specific insertion. The use of alternative pronunciations could be a possible solution for this. Other cases of erroneous segmentation occur in sequences with disfluencies and long creaky effects, in segments with laughter, and in overlapping speech sequences.

5. Impact on acoustic models training

We have retrained a new acoustic model using the modified phone boundaries after applying the above mentioned rules. We have verified that using this second model the WER decreases to 21.5%.

We also have compared the number of correct phone boundaries for a given threshold in the results produced by these two acoustic models, and Figure 3 shows the corresponding results. The graph shows that the phone boundaries produced by the second acoustic model are closer to the manual reference.

6. Impact on Punctuation

In order to analyze the impact of the previous work in the punctuation task, we have conducted two experiments, using the original and modified phone boundaries. We have used a discriminative approach, based on maximum entropy (ME) models, which provide a clean way of expressing and combining different properties of the information. This is specially useful for the punctuation task, given the broad set of available lexical, acoustic and prosodic features. This approach requires all information to be expressed in terms of features, causing the resultant data file to become several times larger than the original one. The classification is straightforward, making it interesting for on-the-fly usage. Experiments described in this paper use the `MegaM` tool [18] for training the maximum entropy models, which uses conjugate gradient and logistic regression. The evaluation is performed using the performance metrics: Precision, Recall and SER (Slot Error Rate) [19]. Only punctuation marks are considered as slots and used by these metrics. Hence, the SER is computed by dividing the number of punctuation errors by the number of punctuation marks in the reference data.

The punctuation experiments here described consider only the two most frequent punctuation marks: *full stop* and *comma*. All the other punctuation marks were converted into one of these two punctuation marks, in accordance with the follow-

	Original			Modified		
	Prec.	Rec.	SER	Prec.	Rec.	SER
Full-stop	68.6	61.6	66.6	68.3	63.9	65.8
Comma	59.5	29.0	90.8	60.0	28.7	90.4
All	64.7	42.6	69.8	64.8	43.4	69.1

Table 1: Punctuation Results.

ing rule: ".": ";", "!", "?", "..." => *full stop*; ",", "-" => *comma*. The training and evaluation tasks are performed using automatic transcripts, produced by our recognition system [15]. The reference punctuation was provided by manual transcripts, initially created by several annotators without linguistic background and recently revised by a linguist expert. Only about 65% of the original punctuation marks were kept in the revised version. The corpus was automatically annotated with part-of-speech information, using MARv [20].

The following features were used for a given word w in the position i of the corpus: w_i, w_{i+1}, $2w_{i-2}$, $2w_{i-1}$, $2w_i$, $2w_{i+1}$, $3w_{i-2}$, $3w_{i-1}$, p_i, p_{i+1}, $2p_{i-2}$, $2p_{i-1}$, $2p_i$, $2p_{i+1}$, $3p_{i-2}$, $3p_{i-1}$, $GenderChgs_1$, $SpeakerChgs_1$, and $TimeGap_1$, where: w_i is the current word, w_{i+1} is the word that follows and $nw_{i\pm x}$ is the n-gram of words that starts x positions after or before the position i; p_i is part-of-speech of the current word, and $np_{i\pm x}$ is the n-gram of part-of-speech of words that starts x positions after or before the position i. $GenderChgs_1$, and $SpeakerChgs_1$ correspond to changes in speaker gender, and speaker clusters; $TimeGap_1$ corresponds to the time period between the current and following word. For the moment, only standard lexical and acoustic features are being used. Nevertheless, this work is a step forward for the use of prosodic features, which already proved useful for this task.

In order to overcome the small amount of speech data we have used an initial punctuation model, trained from written corpora, using only lexical features, to provide the initial weights for the new trains, which use speech transcripts and additional acoustic features. The initial punctuation model was previously trained with two years of written newspaper corpora, containing about 55 million words. Table 1 shows the results for the two punctuation experiments, revealing that the modified data achieves better performances, in terms of SER. Even so, the impact is not as significant as it was expected.

7. Conclusions and future work

Despite the ongoing nature of our work, the results show positive trends. The post-processing rules achieved better results in terms of inter-word pause detection, durations of silent pauses previously detected, and also durations of phones at initial and final sentence-like unit level. Our experiments showed that these improvements had an impact both in terms of acoustic models and punctuation.

This work allows us to tackle more challenging problems, combining prosodic and lexical features for the identification of sentence-like units. It is also a first step towards our goal of adding the identification of interrogatives to our punctuation module.

8. Acknowledgments

The PhD thesis of Helena Moniz is supported by FCT grant SFRH/BD/44671/2008. This work was funded by FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

9. References

[1] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, no. 32, pp. 127–154, 2000.

[2] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, J. Hillard, J. Hirschber, J. Heng, J. G. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Wooters, "Speech segmentation and spoken document processing," *IEEE Signal Processing Magazine*, no. 25, pp. 59–69, 2008.

[3] A. Stolcke, E. Shriberg, T. Bates, M. Ostendorf, D. Hakkani, M. Plauché, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *International Conference on Spoken Language Processing*, 1998, pp. 2247–2250.

[4] R. Carlson, B. Granström, M. Heldner, D. House, B. Megyesi, E. Strangert, and M. Swerts, "Boundaries and groupings - the structuring of speech in different communicative situations: a description of the grog project," in *Fonetik 2002*, 2002, pp. 65–68.

[5] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Speech prosody*, 2002.

[6] D. Wang and S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *ICASSP*, 2004.

[7] J. Vassière, "Language-independent prosodic features," in *Prosody: models and measurements*, A. Cutler and R. Ladd, Eds. Berlin: Springer, 1983, pp. 55–66.

[8] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: a lightweight punctuation annotation system for speech," *Proc. of the ICASSP-98*, pp. 689–692, 1998.

[9] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 35–40.

[10] J. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. Eurospeech*, 2001, pp. 2757–2760.

[11] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proc. of the ISCA Workshop: ASR-2000*, 2000, pp. 228–235.

[12] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transaction on Audio, Speech, and Language Processing*, no. 14, pp. 1526–1540, 2006.

[13] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. of the ICSLP*, 2002, pp. 917 – 920.

[14] B. Favre, D. Hakkani-Tür, and E. Shriberg, "Syntactically-informed modules for comma prediction," in *ICASSP*, 2009.

[15] H. Meinedo, M. Viveiros, and J. Neto, "Evaluation of a live broadcast news subtitling system for portuguese," in *Interspeech*, 2008.

[16] A. Abad and J. Neto, "Incorporating acoustical modelling of phone transitions in an hybrid ann/hmm speech recognizer," in *Interspeech*, 2008.

[17] M. C. Viana, "Para a síntese da entoação do Português," Ph.D. dissertation, University of Lisbon, 1987.

[18] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression," 2004, http://hal3.name/megam/.

[19] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of the DARPA BN Workshop*, 1999.

[20] R. Ribeiro, L. Oliveira, and I. Trancoso, "Using morphossyntactic information in TTS systems: comparing strategies for european portuguese," in *Proc. of PROPOR 2003*, 2003, pp. 26–27.