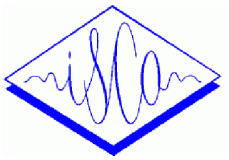
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Integrating Stress Information in Large Vocabulary Continuous Speech Recognition

Bogdan Ludusan, Stefan Ziegler, Guillaume Gravier

CNRS - IRISA, Rennes, France
{bogdan.ludusan, stefan.ziegler, guillaume.gravier}@irisa.fr

Abstract

In this paper we propose a novel method for integrating stress information in the decoding step of a speech recognizer. A multiscale rhythm model was used to determine the stress scores for each syllable, which are further used to reinforce paths during search. Two strategies for integrating the stress were employed: the first one reinforces paths through all the syllables with a value proportional to the their stress score, while the second one enhances paths passing only through stressed syllables, but with a constant value. The former strategy slightly outperforms the later, bringing a relative improvement of more than 2% over the baseline. Furthermore, the stress information proved to be a robust feature, by performing well even for foreign-accented speech.

Index Terms: speech recognition, stress, rhythm

10.21437/Interspeech.2012-507

1. Introduction

A popular approach explored nowadays to improve the performance of speech recognizers is the integration of prosodic information in the recognition process. Among the prosodic components and features investigated, stress seems suitable for speech recognition tasks. This is due to its intrinsic characteristics, stressed syllables exhibiting physical salience, phonemic stability and perceptual distinctiveness [1]. In other terms, their acoustic attributes are more distinguishable, they are less likely to suffer phonological modification and they are recognized better both in clear and noisy conditions. Besides the evidence coming from psycholinguistics studies, support for the use of stress information for recognition can be found also in speech processing works (e.g. [2]).

Despite the general agreement that stressed syllables are more intelligible than their unstressed counterparts, there are only a few recognition systems exploiting this kind of information (e.g. [3, 4, 5]). The systems proposed in [3] and [4] used stress models only for the vowel of the stressed syllable, while the one presented in [5] took into consideration also the consonants of the syllables. The results of these studies were mixed, some of them showing improvements when using stress information [3, 5], while others reporting no performance gain [4].

We propose here a novel approach of combining stress knowledge in the search process by reinforcing paths passing through the stressed syllables. The proposed approach uses a quantitative representation of syllable stress obtained by means of a model of rhythm perception [6]. The model is based on a multiscale filtering mechanism and its corresponding representation, the rhythmogram, indicates the rhythmic grouping of auditory events. The rhythmogram was already applied to speech for the analysis of prosodic structure [7] and of cross-linguistic differences in speech rhythm [8].

This paper is organized as follows: section 2 presents the stress detection procedure and the model on which it is based, the auditory primal sketch. The recognition system employed is described in section 3, while the corpus used in the experiments is presented in section 4. The results obtained are presented next, followed by a discussion on the behaviour and the performances of the proposed system. The paper concludes with some remarks and ideas for future work.

2. Stress Detection

The stress detection procedure is based on a multiscale model of rhythm perception, called the Auditory Primal Sketch [6]. The model aims at identifying the strong acoustic events in the speech signal.

The original model takes into account the effect of the peripheral auditory system on the speech signal by implementing all the steps necessary to model the human ear: the outer and middle ear's transfer functions, the frequency response of the basilar membrane and the response of the inner hair-cells. Next, the response of the auditory nerve is summed across all channels and low-pass filtered with a bank of Gaussian filters having different frequency widths. The peaks of the filters output are combined in a hierarchical representation, with the more prominent events being represented by peaks over more filter widths. This representation is called the rhythmogram.

The auditory nerve response proved to be appropriate for word-level analysis, but the authors argued that for larger time scale events, like phrases, it was more suitable to use the signal energy rather than the nerve response [7].

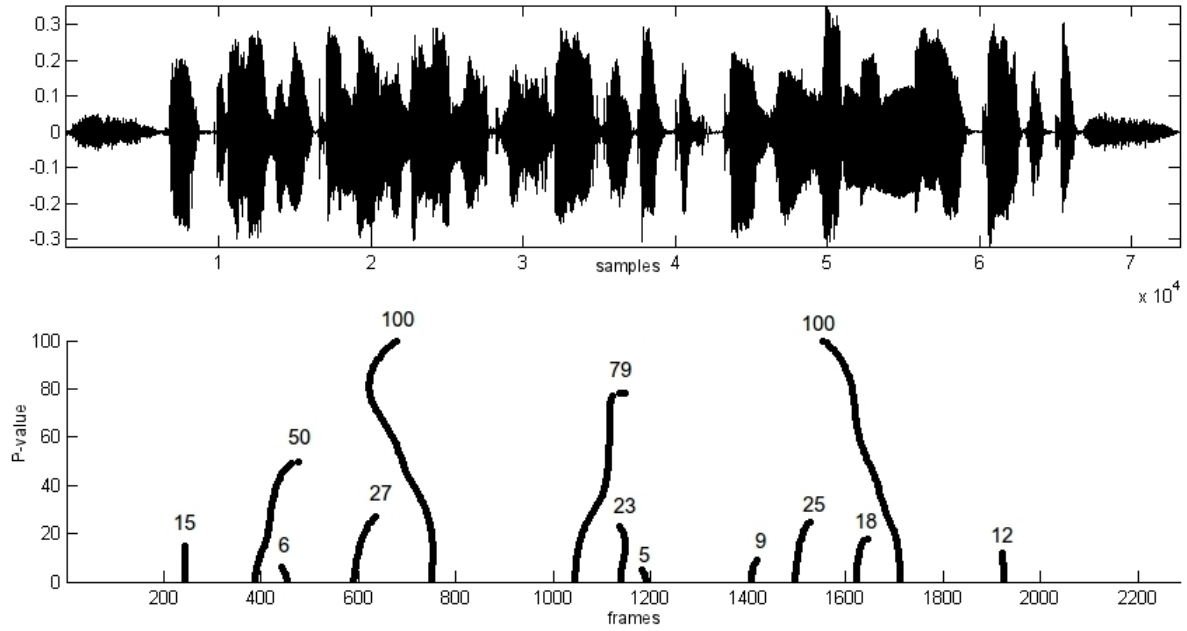


Figure 1: *Waveform and rhythmogram of the sentence "Tous les groupes engagés dans les discussions ont apporté des amendements au texte."*

Thus, we decided to use the energy of the speech signal for computing the rhythmogram in this study. Also, in order to be able to use the information coming from the rhythm model in any automatic process, it has to be quantized. First used in [8], this more compact representation of the rhythmogram consists of a sequence of events and their prominence value (P-value). The P-values were obtained by summing all peak values of a given event across all filter widths.

Figure 1 illustrates the waveform of a sentence from the corpus used in the experiments (upper panel) along with the obtained rhythmogram (lower panel). The value above each event symbolizes its corresponding P-value.

After computing the rhythmogram of each phrase, the stress value for each syllable was determined in a similar manner to the one used for computing the stress accent score in [9]:

- search inside the boundaries of each syllable and determine its corresponding rhythmogram events
- in case of multiple events corresponding to the same syllable, take only the event with the highest P-value
- the stress score is equal to the P-value of the event
- if no event was found within the boundaries of a syllable, set its stress score to 0

Being a preliminary study, the syllable boundaries were obtained by combining phone-level forced aligned data with the French syllabification rules proposed by Dell [10].

3. Recognition System

The recognizer used in the experiments, Irene, is a two-pass recognition system developed at IRISA. The first pass generates a word graph, while the second pass performs the rescoring of this graph using more complex models.

The first pass is based on the algorithm proposed by Ortmanns and Ney [11] and it implements a beam-search strategy for decoding. The acoustic models employed are word-internal triphones with 4,019 distinct states and 32 Gaussians per state and word trigrams are used as language model. The rescoring pass uses 4-grams as language model and cross-word triphone models with 6,000 states and 32 Gaussians each. No speaker adaptation or morpho-syntactic information were used in the second pass.

The search strategy used by the recognizer aims at maximizing the following equation:

$$Q(j,t) = \max_i Q(i,t-1) + log(a_{ij}) + log(b_j(y_t)) \quad (1)$$

where $Q(j,t)$ is the score of the path in state j at time t, $log(a_{ij})$ the transition probability between states i and j and $log(b_j(y_t))$ the observation probability of y_t when in state j.

In order to accommodate for the stress information, the previous equation was changed by adding a reinforcement factor:

$$Q(j,t) = \max_i Q(i,t-1) + log(a_{ij}) + log(b_j(y_t)) + str(t) \cdot R \quad (2)$$

where $str(t)$ represents the stress score of the syllable at time t and R is the maximum reinforcement factor. The stress score varies between 0, meaning least stressed, and 1, representing most stressed. The parameter R is used for weighting the stress score and its value was obtained by optimizing the recognition performance on a development set.

The reasoning for making this implementation choice seems logical considering the results of studies showing that stressed syllables are better recognized than non-stressed syllables [1, 2]. Thus, by reinforcing the paths passing through the stressed syllables, we effectively give those syllables a higher weight in the search process.

4. Materials

The data employed for the experiments was taken from the French ESTER2 evaluation campaign corpus [12]. The corpus consists of mainly radio broadcast news, with a small part of it containing more spontaneous material, like debates. While all its recordings are in French, the corpus contains also news from African radio stations, which exhibit strong accents. Approximately 3.5 hours from the ESTER2 development set were used for tuning the parameters, while 4.5 hours from the ESTER2 test set were used for the experiments.

5. Results and Discussion

Two factors were considered when choosing the experimental settings. The first one concerned how the stress information should be integrated in the decoding process: either by adding to the path score a value proportional to the stress score of each syllable, or by reinforcing the path score with a constant value, but only for the stressed syllables. Here, a syllable was considered to be stressed if its stress score was greater than 0.5. The second factor addressed the issue of where, inside the recognition chain, the stress information should be added. A straightforward choice would be to add it to the first pass, because a better word graph before the rescoring pass would translate into increased performance. In case such an improvement was found after a stress-enhanced first pass, we were interested in knowing whether the stress would bring new information in the second pass, even if more complex acoustic models were being used.

A total of four tests were performed, by using either a stress-proportional or a constant value for reinforcing paths, and by integrating the stress information either only in the first pass or in both passes. Table 1 presents the Word Error Rate (WER) obtained on the development set with the baseline, the Irene recognition system, and the recognizer integrating stress information in the two previously explained configurations. The proportional reinforcing strategy improved recognition both when applied only in the first pass, as well as in both passes, while the strategy using a constant reinforcement posted a low gain.

Reinforcement strategy	Baseline	Stress	
		1st only	1st & 2nd
proportional	35.6	35.3	35.1
constant		35.4	35.4

Table 1: *WER (in %) obtained on the development set.*

The results obtained on the test set are presented in Table 2. When applying the proportional stress-enhancing scheme in the first step the performance was improved by 1.3% relative value, while its use in both passes brought a relative WER reduction of 2.1%. A Wilcoxon signed rank test was applied to the results obtained and it found that the differences are significant at the 1% and 0.1% level, respectively. Although the constant reinforcing strategy had a poor performance on the development set, it actually brought an improvement on the test set. The same statistical test showed a significant difference at the 0.1% level when integrating stress in both passes, with respect to the baseline.

Reinforcement strategy	Baseline	Stress	
		1st only	1st & 2nd
proportional	23.5	23.2	23.0
constant		23.4	23.1

Table 2: *WER (in %) obtained on the test set.*

The better strategy among the two, the proportional one, was further used in experiments to provide a more detailed discussion of the behaviour and of the performances obtained by integrating stress information in the decoding step. First, we took a look at the WER on the development set for different values of R. Figure 2 shows the WER of the baseline (solid line), the results by varying R in the first pass (dashed line) and by varying it in the second pass (dotted line). One can see that the performance does not worsen even for high values of R, especially in the first pass, meaning that stress offers a robust source of information.

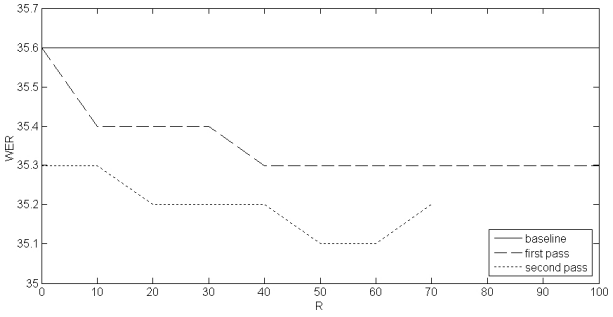


Figure 2: *WER (in %) on the development set, for different values of the R parameter.*

A measure which can quantify the gain obtained by integrating stress information in the first pass is the Word Graph Error Rate (WGER). It is computed by determining the best sentence through the graph that matches the reference transcription and it offers a lower bound of the WER. As seen in Table 3, an absolute improvement of 0.6% in WGER was obtained by using stress information in the first step. Because this improvement is higher than the 0.3% decrease in WER after the same step, a new approach in the second pass was needed in order to fully take advantage of the gain in WGER. It can be seen from the results presented in Table 2 that the integration of stress information also in the rescoring pass does indeed achieve this goal.

	Baseline	Proposed
WGER	12.7	12.1

Table 3: *WGER (in %) after the first pass.*

As mentioned in Section 4, the corpus contains also recordings exhibiting strong accents. Because the acoustic models were mostly trained on standard French broadcast news, we decided to examine how the stress-enhanced recognizer behaves on this type of speech. It can be observed in Table 4 that the baseline performs worse on the strong-accented data, with a relative WER more than 50% higher than for non-accented speech. The stress-based decoding performs well even in this difficult conditions, bringing a higher relative improvement when used only in the first pass with respect to the non-accented speech (1.9% vs. 1.5%) and the same level of improvement when used in both passes (2.5%).

System	Stress	Accent		Overall
		Normal	Strong	
Baseline	-	20.0	31.5	23.5
Proposed	1st only	19.7	30.9	23.2
	1st & 2nd	19.5	30.7	23.0

Table 4: *WER (in %) on the test set based on the type of accent present.*

6. Conclusions

We investigated in this paper the usefulness of incorporating stress information in a large vocabulary speech recognition task. Differently from previous speech recognition studies using stress information, no prior stress knowledge is stored in the lexicon, nor was it used during the training stage. The stress information was obtained by means of a model of rhythm perception, which aims at finding strong acoustic events inside the speech signal. A small but statistically significant improvement was obtained (0.5% absolute value). The improvements were found to be due to differences in the behaviour of the systems during hypothesis propagation, the addition of stress information helping to prune away some of the unreliable hypotheses. Also, it appears that the stress score used in this paper is a robust feature, bringing improvements when both simpler and more complex acoustic models were used and both in the presence of normal and strong-accented speech. Its robustness was further proved by the large interval of values that the parameter R can take without decreasing the recognition performance.

As a future step, we plan to see how the recognition process behaves when automatically segmented syllables are used, or when only the neighbourhood around a rhythmogram event, and not the entire syllable, is reinforced. While in this study the role of the stress was to enhance paths equally inside a syllable, it would be interesting to use different reinforcement values inside the same syllable by weighting them based on each frame's acoustic likelihood.

7. Acknowledgements

This work was partially supported by the Agence Nationale de la Recherche through the ASH project.

8. References

[1] Mattys, S., "The use of time during lexical processing and segmentation: A review", Psychonomic Bulletin and Review , 4: 310–329, 1997.

[2] Jenkin, K. and Scordilis, M., "Development and comparison of three syllable stress classifiers", In Proc. of ICSLP-96, 733–736, 1996.

[3] Wang, C. and Seneff, S., "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain", In Proc. of EUROSPEECH-2001, 2761–2764, 2001.

[4] van den Heuvel, H., van Kuijk, D. and Boves, L., "Modelling lexical stress in continuous speech recognition", Speech Communication, 40: 335–350, 2003.

[5] van Dalen, R., Wiggers, P. and Rothkrantz, L., "Lexical stress in continuous speech recognition", In Proc. of INTERSPEECH-2006, 2382–2385, 2006.

[6] Todd, N., "The auditory primal sketch: a multi-scale model of rhythm grouping", Journal of New Music Research, 23: 25–70, 1994.

[7] Todd, N. and Brown, G., "Visualization of rhythm, time and metre", Artificial Intelligence Review, 10: 253–273, 1996.

[8] Lee, C. and Todd, N., "Towards an auditory account of speech rhythm: application of a model of the auditory primal sketch to two multi-language corpora", Cognition, 93: 225–254, 2004.

[9] Ludusan, B., Origlia, A. and Cutugno, F., "On the use of the rhythmogram for automatic syllabic prominence detection", In Proc. of INTERSPEECH-2011, 2413-2416, 2011.

[10] Dell, F., "Consonant clusters and phonological syllables in French", Lingua, 95: 5–26, 1995.

[11] Ortmanns, S. and Ney, H., "A word graph algorithm for large vocabulary continuous speech recognition", Computer Speech and Language, 11: 43–72, 1997.

[12] Galliano, S., Gravier, G. and Chaubard, L., "The ESTER 2 evaluation campaign for the rich transcription of French broadcasts", In Proc. of INTERSPEECH-2009, 1149–1152, 2009.