Is Syllable Stress Information Robust for ASR in Adverse Conditions?

Bogdan Ludusan, Stefan Ziegler, Guillaume Gravier

CNRS - IRISA Rennes, France

bogdan.ludusan@ens.fr, {stefan.ziegler, guillaume.gravier}@irisa.fr

Abstract

This paper presents a study on the robustness of stress information for automatic speech recognition in the presence of noise. The syllable stress, extracted from the speech signal, was integrated in the recognition process by means of a previously proposed decoding method. Experiments were conducted for several signal-to-noise ratio conditions and the results show that stress information is robust in the presence of medium to low noise. This was found to be true both when syllable boundary information was used for stress detection and when this information was not available. Furthermore, the obtained relative improvement increased with a decrease in signal quality, indicating that the stressed parts of the signal can be considered islands of reliability.

Index Terms: speech recognition, prosody, syllable stress, noise

1. Introduction

Prosodic information has already been used successfully in large vocabulary speech recognition (e.g. [1, 2]). Prosodic information (duration, pauses, F0, etc) was integrated in automatic speech recognition (ASR) systems both at the acoustic and language model level, being posited that prosodic features are robust to noise and unaffected by channel condition [1].

Among the major prosody components, stress seems to present several characteristics which are particularly helpful for speech recognition tasks under different conditions. Studies examining the role of stressed syllables showed that they provide salience in terms of their acoustic attributes, they are less likely to suffer phonological modification or to be misinterpreted and they are detected more consistently than unstressed syllables in *noisy* environments [3].

Indirect evidence supports the relevance of stress information in noisy conditions, not only for humans, as shown by psycholinguistics studies [4] but also for machines [5]. A human perception in noise study [4] examined the intelligibility of speech at very low signal-to-noise ratios (SNR), by either masking or unmasking the stressed syllables. It showed that the SNR required to identify the consonants of the unstressed syllables increased when the stressed syllables were masked and it decreased when the stressed syllables were unmasked. The author concluded that the listener relies on prosody to achieve robust speech understanding and that the information in the stressed syllables helps predict the neighbouring unstressed syllables. Further evidence can be found in an ASR study investigating the effects of phonetic information reduction on recognition performance [5]. When phone identity was substituted with manner feature, performance dropped in both clean and noisy conditions, but no significant difference was observed when phone identity only inside unstressed syllables was replaced by manner information. This suggests that the information carried by the stressed syllables has a higher importance for speech recognition in noisy environments than the one present in non-stressed syllables.

Based on the fact that humans exploit stress information not only under normal acoustic conditions, but also in the presence of noise, and that this information is salient for them, we are interested in investigating whether adding stress information to ASR in adverse conditions would be useful. One would expect it, as stressed syllables display higher energy than non-stressed syllable and in noisy environments they would exhibit a higher SNR than the neighbouring syllables, thus helping to their recognition.

Stress information has been used before in several speech recognition systems [6, 7, 8, 9, 10]. The systems employed different methods to add this new information to the recognition process: at the lexicon level [7, 8], as a separate model in the decoding process [6, 9], or to guide paths during search [10]. Most of these studies reported statistically significant improvements when including stress, but the role of stress was examined exclusively under normal acoustic conditions. To our knowledge, there is only one study in the literature which reported results for speech under adverse conditions [10]. In that study, experiments both on normal and strong accented speech were performed and the same level of improvement was obtained for the two types of speech, when stress information was used. A close analysis of the system showed that the integration of stress information improved speech recognition due to its interaction with the pruning process, by helping prune away some of the wrong hypotheses.

The current paper builds upon a previous mentioned study [10] and aims at the following two aspects: to enlarge the investigation of stress robustness for ASR in adverse condition to a new case (additive noise) and to explore the behaviour of the system when there is no syllable boundary information available for the computation of the stress score. Because we wanted to examine only the effect of stress on the recognition performance, the same system was used in the clean and noisy condition, with no speech enhancing pre-processing. Further details on the recognition system employed and on the stress detection procedure can be found in section 2. For the recognition experiments presented in section 3, we used white and pink noise at various SNRs and we tested two conditions based on whether syllable boundary information was available for stress detection or not. We have chosen to use coloured noise in this study as a starting point in our research, an initial test, but we envisage our future work to include more complex types of noise.

2. System Presentation

The system used in this paper is composed of two components: a stress detection procedure and a speech recognizer. The first component computes syllable-level stress scores which are sub-

10.21437/SpeechProsody.2014-29

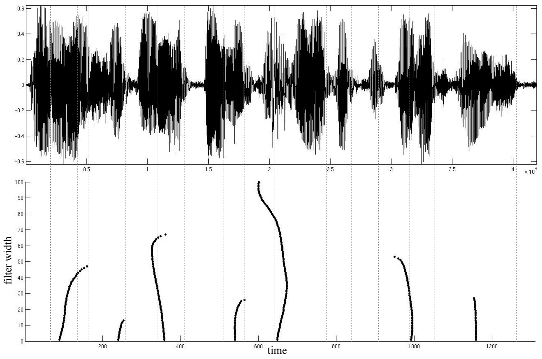

Figure 1: Speech waveform and corresponding rhythmogram.

sequently integrated in the search process of the ASR system.

2.1. Stress Detection

For the computation of the stress score, an unsupervised method based on the auditory primal sketch (APS) [11] was used. The APS is a model of rhythm perception which seeks to identify strong acoustic events present in speech, in a similar way to edge detection in vision theory [12]. In order to accomplish this, the speech signal is filtered with a bank of Gaussian filters with different filter widths. The peaks of the obtained functions are stacked onto each other, with the peaks of the minimum filter width at the bottom. The stacking of the maxima forms contiguous lines which we call "events" (as exemplified in Figure 3). Thus, a hierarchical representation of speech, called a rhythmogram, is obtained. The rhythmogram contains the time on the abscissa and the event height on the ordinate and can be condensed into a two-line vector, containing the time instants of the events on the first line and their corresponding values on the second line.

The procedure used for obtaining the rhythmogram, as well as the values of the parameters needed for its computation, are the one proposed in a previous study [13]. It consists of the following steps:

1. resample at 500 Hz,
2. perform full wave rectification,
3. take the cubic root, to model the ears' loudness function,
4. apply one hundred logarithmically-distanced Gaussian filters,
5. stack the maxima of the obtained function in a 2-D representation, with time on the x-axis and filter width on the y-axis.

Once the rhythmogram is computed, information about syllable boundaries is needed in order to obtain a syllable-level stress score. Then, a search within each syllable is performed and the value of the highest event is taken as the stress score. For the experiments conducted in section 3.2 we had this information available, while the experiments presented in section 3.3 made use of an approximation. Further details will be given in the respective sections.

Figure 1 illustrates the waveform of a sentence from the corpus used in the experiments along with its corresponding rhythmogram. Each point of the rhytmogram is associated to a time instant (t) and a filter width (i) and it represents a peak at time t in the function obtained with the ith filter. For this particular example we can observe that for the minimum filter width (bottom of lower panel) the function had 7 maxima, for the 40th filter 4 maxima, while applying the filter with the highest width returned a function with one maximum (top part of lower panel). By plotting these points in space, we obtain the contiguous lines observed in the figure (the events). In order to compute the stress score we determine to which syllable each event belongs, based on the start time of the event, obtained with the lowest filter width. The stress score will be equal to the number of points which form the event, in this case a stress score of 0 is obtained for the first syllable, a stress score of 47 for the second one, etc.

2.2. Speech Recognition System

For all the experiments we used a two-pass recognition system [10]. The recognizer, produced a word graph after the first pass, graph which will be rescored using more complex acoustic models in the second pass. It uses in the first pass word-internal triphone acoustic models with 4,019 distinct states and 32 Gaussians per state and word trigrams as language model. The rescoring pass has 4-grams as language model and cross-word triphone models with 6,000 states and 32 Gaussians each.

In order to integrate the stress information in the recognition system, the Viterbi search was modified as shown in Equation 1.

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

The first three terms are also present in the classical Viterbi decoding: $Q(j,t)$, the score of the path up to 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. The last term is a product between $str(t)$, the stress score of the syllable at time t (taking values between 0 and 1), and R, a weighting factor which represents the contribution of the stress information to the decoding process. The value of R was determined by optimizing the recognition performance on the development set.

Thus, by adding the new term in the search equation, the decoding procedure reinforces all the phonemes belonging to the stressed syllables with a value proportional to the syllable stress score. This implementation choice was made based on the fact that stressed syllables are more stable and are distinguished better [3] and, by giving them a higher weight in the search process, improvements can be obtained [10].

3. Experiments

Speech recognition experiments were conducted on a corpus of broadcast news to which coloured noise, at different SNRs, was added. We investigated the role of syllable boundary information for the computation of the stress score as well as the effect of optimizing the stress weighting factor for each noise level. For each of the experiments, the recognition performance was evaluated by considering the reduction brought in terms of word error rate.

3.1. Materials

The materials used for the experiments presented here are the files of the ESTER 2 evaluation campaign corpus [14]. The corpus consists of mainly broadcast news recordings from French radio stations, although it also contains some more spontaneous radio shows as well as recordings from French-speaking African radio stations which exhibit strong accents.

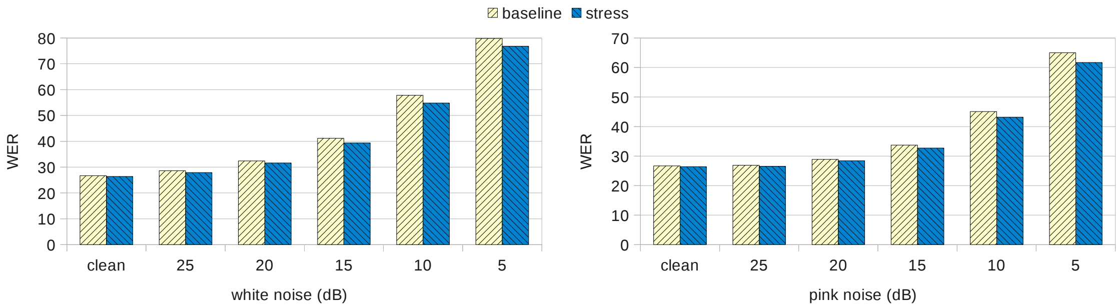

Figure 2: The WER obtained for various SNRs of additive white noise (left panel) or additive pink noise (right panel).

The entire training set, approximately 180 hours of data, was used for the estimation of the acoustic models. Its corresponding transcriptions, along with articles from newspapers, were employed in the computation of the language model. The rest of the corpus forms the development set, used for tuning the system parameters, and the test set, on which the performance of the system was evaluated. These subsets contain circa 6 and 7 hours of recordings, respectively.

For the experiments, two types of additive noise were added to the speech files: white noise and pink noise, ranging from 25 to 5 dB SNR. The white noise was generated using the MATLAB function *wgn*, while the pink noise was obtained by filtering the previously generated white noise [15]. For each SNR, the gain of the filter was determined in order to obtain the desired SNR after the filtering operation. The noise was added only to the files belonging to the development and evaluation sets, the acoustic models having been previously trained with the original training set files.

3.2. Experiment 1

As mentioned in section 2.2, in the first experiment we had knowledge of syllable boundaries for computing the syllable stress score. The boundaries were determined by force aligning the data at the phoneme level and then applying French syllabification rules [16]. Although the quality of the syllabification was not as good as when manual syllables would have been used, this was the best available option for computing a syllable-level stress score. The search for the parameter R in Equation 1 was performed only on the clean data and the obtained value was used in all experiments, regardless of the SNR. The stress score was determined at each step from the noisy data.

The recognition results for various levels of noise are presented in Figure 2: for white noise (WN) in the left panel and for pink noise (PN) in the right panel, with the reported measure being the word error rate (WER). The results obtained with the baseline system are represented by the lighter coloured columns, while those of the recognizer employing stress knowledge are represented by the darker coloured columns. The clean condition is illustrated in both panels for an easier comparison. It can be observed that the importance of stress information in the recognition process increases with the decrease in SNR.

$$WER_{rel} = \frac{WER_{stress} - WER_{base}}{WER_{base}} \quad (2)$$

Next, we used the relative WER (see Equation 2) to compare the performance at different SNRs of the speech signal. As can be seen in Figure 2, the results of the system using stress information were always better than those of the baseline. Thus, we show in Table 2 the values of $abs(WER_{rel})$, which represent the relative improvement (in %), at each SNR, with respect to the baseline. A Wilcoxon signed rank test was used to determine the statistical significance of the results. All the differences were found to be *significant* at the $p < 0.001$ level, except for PN25 ($p < 0.01$).

Noise	Clean	25 dB	20 dB	15 dB	10 dB	5 dB
WN	1.1	2.5	2.5	4.4	5.2	3.8
PN		1.5	1.7	3.0	4.2	5.1

Table 1: *Relative improvement (in %) when syllable boundaries are known.*

The same tendency is observed for both types of noise: the relative improvement increases with the decrease in signal quality. This suggests that the information in the stressed parts tends to have a higher weight on the recognition process as conditions deteriorate. Also, because we were using a value for R determined on clean data it also proves the robustness of the information added. The results obtained in this experiment can be considered as an upper bound for the improvement brought by adding stress information, as it uses almost ideal syllable boundary information.

3.3. Experiment 2

In the second experiment no syllable boundary information was used. Instead we define a time interval around each rhythmogram event which will act as a “pseudo-syllable”, i.e. the whole region will be considered as one entity and it will be assigned a stress score equal to the height of its corresponding rhythmogram event. The gaps between the obtained regions will be assigned a zero stress score and, thus, they will have no effect on the decoding process. Similarly to the first experiment, the value of R was determined on the clean development set, while the stress score was obtained from the noisy data.

The “pseudo-syllable” approach is illustrated in Figure 3. The curly braces under each event correspond to the size of the region considered, while the intervals delimited by dashed lines and labeled PS_n represent the final entities. Their size might be lower than the chosen region size, due to events being too close to each other, or too close to the beginning or the end of an utterance. In order to choose the size of the region we took a look at the average size of the dev set syllables having non-zero stress scores. The value obtained, 197 ms, can be compared to the 141 ms average length of syllables having stress scores equal to zero. By rounding the 197 ms to the closest odd number of frames (the central frame and an equal number of frames on

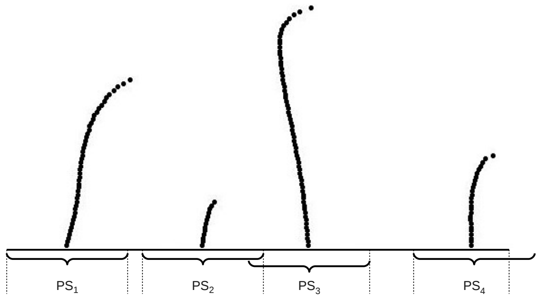

Figure 3: The approach used for hypothesizing syllables, when syllable boundaries are unknown.

each side) we get regions 19 frames wide. This value was used in the present experiment.

Table 2 shows the results given by the recognizer when the previously explained approach is employed. All differences except for PN25 and WN20, were found to be statistically *significant* (WN25 $p < 0.05$, WN15-WN5 $p < 0.01$, PN20 $p < 0.01$, PN15 $p < 0.05$, PN10-PN5 $p < 0.001$). Although the performance increase obtained is not as important as in in the case when syllable boundaries were known, a similar trend in the results can be observed.

Noise	Clean	25 dB	20 dB	15 dB	10 dB	5 dB
WN	0.8	0.7	0.3	1.2	1.4	1.0
PN		0	0.7	0.6	1.1	1.9

Table 2: *Relative improvement (in %) when syllable boundaries are unknown.*

3.4. Experiment 3

As a final test, we examined the effect of parameter optimization on the recognition performance. The experiment was run only on the data containing additive white noise and an optimum value of the R parameter was obtained for each SNR, on the dev set. Both conditions used in the previous two experiments, with or without knowledge of syllable boundary information, were considered and the results obtained are presented in Table 3.

Condition	25 dB	20 dB	15 dB	10 dB	5 dB
Exp 1	2.5	2.5	3.2	5.5	5.0
Exp 2	1.1	1.9	3.4	3.3	0.5

Table 3: *Relative improvement (in %) for an optimized value of the R parameter, in the case of white noise.*

Comparing the results in the first row with those illustrated in section 3.2, one can see that they are quite *similar*, the only case where a significant improvement was obtained is for the 5 dB level ($p < 0.001$). Furthermore, the performance advantage obtained in Experiment 1 for 15 dB is statistically significant ($p < 0.01$). This shows that, when the syllable boundaries are known, the values of R obtained in clean conditions give a good performance also for noisy speech. And, by not needing to optimize R for each SNR, we avoid having to estimate the noise level prior to the actual recognition process.

For the conditions described in Experiment 2, the optimization of the parameter R gives instead *significant* improvements for all noise levels (WN25, WN5 $p < 0.05$, WN20-WN10 $p < 0.001$). An interesting results was obtained for the 5 dB WN condition: although the performance was lower in all the experiments conducted, here the difference was the highest. This might be due to the length of the region considered for our syllable approximation. In case of low SNR it would probably more appropriate to consider smaller regions. While this will decrease the effect that stress has on recognition, by reinforcing a smaller area it is more likely that this area will fall inside the boundaries of the actual syllable and it will not introduce any other errors.

4. Conclusions

In this study we investigated the robustness of stress information in the recognition process, when speech is corrupted by additive noise. To our knowledge this is the first study in the literature aimed at investigating this issue. Using white and pink noise in the experiments conducted, we have observed the same behaviour in both cases: for medium to low SNRs, higher relative improvements are obtained with the increase in the noise level, when stress knowledge is integrated into the recognizer. These results support the view that stressed syllables represent the reliable regions of the speech signal and that the information they carry is important for speech recognition. Besides agreeing with the role given to stress by psycholinguistic studies [3, 4] as well as to the indirect evidence coming from other ASR studies [5] the results of this investigation also encourage the use of such information in speech recognition.

Stress information is robust in the presence of additive noise, especially when syllable boundaries are known for the computation of the stress score. This finding is supported by the small difference observed when the R parameter was optimized for each noise level, compared to the case when the value for R obtained on the clean data was used in the experiments. While improvements are obtained also when a syllable approximation is used instead of the actual syllables, they are significantly lower and depend more on the value of the R parameter, for different noise levels. This might suggest that the approximation used is not suitable for calculating the stress score and that a new approach should be sought. A possible alternative would be to hypothesize syllable boundaries based on the transcription obtained after the first pass of the recognizer. Unfortunately this approach would then limit the use of stress information to the second step only and it was shown in [10] that a big part of the improvement brought by stress information was due to its use in the first pass.

In this work we examined the role of stress information only in the case of coloured noise, but, in the future, we plan to extend the study to also include speech in the presence of competing talkers. Further lines of research to follow include the search for a better syllable approximation or the use of stress only in the rescoring step, as syllable information can be extracted from the word graph produced by the first pass. Also, we used a 19-frame representation in the present work, but a search for the optimum size of the region to be considered might give better results.

5. Acknowledgements

This work was partially funded by the Agence Nationale de la Recherche, through the ASH project. The first author is now affiliated with LSCP - EHESS/ENS/CNRS, Paris.

6. References

[1] D. Vergyri, A. Stolcke, V. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. of IEEE ICASSP 2003*, 2003, pp. 208–211.

[2] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, T. Yoon, and S. Chavarria, "Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus," *Speech Communication*, vol. 46, pp. 418–439, 2005.

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

[4] P. Divenyi, "Humans glimpse, too, not only machines (hommage à Martin Cooke)," in *Forum Acusticum 2005*, 2005, pp. 1533–1538.

[5] E. Fosler-Lussier, A. Rytting, and S. Srinivasan, "Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on ASR performance," in *Proc. of INTERSPEECH-2005*, 2005, pp. 1249–1252.

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

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

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

[9] S. Ananthakrishnan and S. Narayanan, "Prosody-enriched lattices for improved syllable recognition," in *Proc. of INTERSPEECH-2007*, 2007, pp. 1813–1816.

[10] B. Ludusan, S. Ziegler, and G. Gravier, "Integrating stress information in large vocabulary continuous speech recognition," in *Proc. of INTERSPEECH-2012*, 2012.

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

[12] D. Marr, *Vision*. New York: Freeman Education, 1982.

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

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

[15] J. Smith, *Spectral Audio Signal Processing*. W3K Publishing, 2011, ISBN: 978-0-9745607-3-1.

[16] F. Dell, "Consonant clusters and phonological syllables in French," *Lingua*, vol. 95, pp. 5–26, 1995.