LANGUAGE-INDEPENDENT AUTOMATIC SYLLABLE SEGMENTATION USING BROAD PHONETIC CLASS INFORMATION

Bogdan Ludusan, Emmanuel Dupoux

Laboratoire de Sciences Cognitives et Psycholinguistique
EHESS / École Normale Supérieure, PSL Research University / CNRS, Paris, France

ABSTRACT

We propose in this paper a language-independent method for syllable segmentation. The method is based on the Sonority Sequencing Principle, by which the sonority inside a syllable increases from its boundaries towards the syllabic nucleus. The sonority function employed was derived from the posterior probabilities of a broad phonetic class recognizer, trained with data coming from an open-source corpus of English stories. We tested our approach on English, Spanish and Catalan and compared the results obtained to those given by an energy-based system. The proposed method outperformed the energy-based system on all three languages, showing a good generalizability to the two unseen languages. We conclude with a discussion of the implications of this work for under-resourced languages.

***Index Terms*—** syllable segmentation, sonority, broad phonetic class, posterior probabilities

1. INTRODUCTION

The syllable is the smallest prosodic units and it plays an important role in the description of all prosodic phenomena. Similarly to other speech annotations, syllable segmentation is a time consuming task and automation of this process is desired, in order to be able to process large datasets. Information about syllables is useful not only for phonetic analysis of corpora, but also in speech technology applications, having been used for speech rate estimation [1], or the automatic detection of prosodic events (e.g. acoustic prominence [2], prosodic boundaries [3]).

A popular automatic syllable segmentation method is based on the energy of the speech signal (e.g. [4, 5]). It offers the advantage of being language-independent, but requires the setting of a number of parameters, and its performance is sensitive to recording conditions. Another approach for language-independent automatic syllable segmentation can employ knowledge from the phonological theory. In linguistics, sounds can be grouped in classes, based on various criteria. One such criterion is the manner of articulation and the division of the phonetic space based on this criterion will be called throughout the paper as broad phonetic classes. Each broad phonetic class has a different level of sonority, from obstruents, with a low sonority, to vowels, represented by a high sonority. For segmentation, one can apply the Sonority Sequencing Principle (SSP) [6], which states that the sonority inside a syllable increases towards the nucleus and and then decreases again towards the left edge.

We propose a system based on the SSP, which uses a speech recognizer, trained on an open-source corpus of English, to obtain the probabilities of each broad phonetic class. These probabilities are then combined with the sonority values of each class to derive an overall sonority function and syllable nuclei and boundaries are placed in correspondence to the maxima and minima of this function. Similar methods have been proposed for speech-based nucleus detection [7] and syllable segmentation [8]. A broad phonetic class recognizer was used to obtain the vocalic nuclei of syllable in order to estimate the speech rate [7]. Automatic syllable segmentation was performed in [8] by force aligning the speech signal, then taking the sonority values of the obtained phonemes and placing syllable boundaries in correspondence to the minima of this function. Differently from these approaches, we do not use the recognizer to produce a sequence of phonemes/phonetic classes, but to determine the posterior probability of each frame and we derive from it a continuous sonority function. Thus, we are not limited only to the class decision taken by the recognizer [7], but can take into account the contribution of all the classes. Also, by using phonetic recognition, not forced alignment [8], we can apply it to languages that do not have trained acoustic models.

The paper is further structured: Section 2 presents in detail the two components of the syllable segmentation system, namely the speech recognizer and the nuclei and boundary placement function. The datsets used in the experiments and the results obtained are detailed in Section 3. The paper concludes with a discussion on the performance of the system and its possible use for under-resourced languages.

The research leading to these results was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON). It was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the École des Neurosciences de Paris, and the Région Île-de-France (DIM cerveau et pensée).

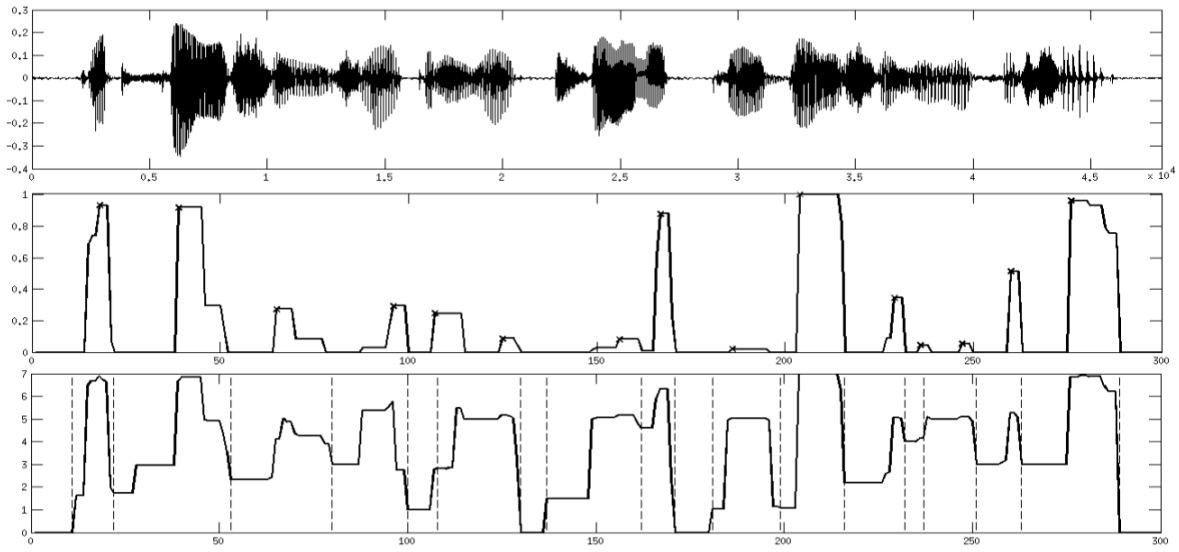

Fig. 1. Waveform of the phrase "It functions like an electronic probation officer." (upper panel) and corresponding nucleus sonority (middle panel) and total sonority (lower panel). The position of the found nuclei is marked with an X sign in the middle panel, while the obtained syllable boundaries are marked by a dashed line in the lower panel.

2. METHODS

The segmentation procedure is performed in two steps: First, a speech recognizer is used to decode the input sequence into broad phonetic classes. Second, we use the posterior probabilities given by the recognizer to derive two functions: a nucleus sonority and a total sonority function, based on which the syllable nuclei and boundaries will be placed. We will describe in detail the two components of the system.

We use here a sonority scale similar to the one proposed by Clements [6] (vowels>glides>liquids>nasals>obstruents), by further dividing the obstruent class in three sub-classes (fricatives>affricates>plosives), for a better modelling of the obstruent phonemes. Thus, we use a 7-steps sonority scale, with the value 7 corresponding to the vowel class and plosives having a sonority value of 1. The silence intervals were given a sonority value equal to 0.

2.1. Broad phonetic class recognizer

A broad phonetic class recognizer was employed in the first step to obtain the posterior probabilities of the 8 classes (7 broad phonetic classes + silence) defined in this study. It was built using the Kaldi toolbox [9] and was trained with recordings from Librispeech [10], an open-source corpus of English stories, mainly used for automatic speech recognition. The Librispeech subset employed for the training of the acoustic models was the train-clean-100, containing 100.6 hours of recordings coming from 251 speakers (125 females, 126 males). The values of the training parameters are the ones given by the Kaldi Librispeech recipe. A unigram language model with flat probabilities was chosen, in order not to bias our phonetic recognizer to the phonotactics of English. 13 Mel frequency cepstrum coefficients were extracted, along with their deltas and double deltas, from a 25 ms analysis window, every 10 ms. In this study we investigated how the use of different acoustic models would impact the segmentation performance, so three acoustic models were tested:

- monophone model (*mono*)
- triphone model using Linear Discriminant Analysis (LDA) transforms and Maximum Likelihood Linear Transform (MLLT) estimation (*triA*)
- triphone model with LDA+MLLT and speaker adaptive training (*triB*)

2.2. Syllable segmentation

Once the posterior probabilities for the eight classes have been extracted from the speech recognizer, we used them to compute two sonority functions: a nucleus sonority and a total sonority. The former is used for the detection of the syllabic nuclei, while the latter for the placement of the syllable boundaries. For each analysis frame, the total sonority is defined as follows:

$$totSon_k = \sum_{i=0}^{7} prob_{ki} * sonority_i$$

where $prob_{ki}$ represents the posterior probability of class i, at frame k, while $sonority_i$ is the sonority of class i, as given by the sonority scale introduced in section 2 (silence=0, plosives=1 and so on).

The nucleus sonority is computed in a similar manner, by reducing the sum over all classes to the only one class, that representing the vowels. Thus, its value will be directly proportional to the posterior probability of the vowel class.

After the two sonority functions are computed, they are processed to remove some unwanted phenomena, since no smoothing is used on the sonority functions. In the case of the nucleus sonority, any spurious one-frame maxima or minima that change the monotonicity of the function (on upwards or downwards slopes) were removed, as they might introduce additional local peaks. For the total sonority function we marked all frames having a sonority lower than 1 (equivalent to plosives) as being silence frames and remove any one-frame long silences. We then search the nucleus sonority function for local maxima that do not fall inside a silence interval (defined by a total sonority value of 0) and consider them as being syllable nuclei candidates. Syllable boundaries are afterwards placed in correspondence to the local minima of the total sonority, between each two syllable nuclei candidates. As a final step, all syllables found to be shorter than 25 ms (the length of an analysis frame) are removed.

3. EXPERIMENTS

We present here the experimental setting used in this study, by introducing the corpora on which the proposed approach was tested and the evaluation measures employed, followed by the obtained results.

3.1. Materials

Three languages were used for the experiments: English, the language on which the acoustic models were trained on, and two unseen languages, Catalan and Spanish. We hoped that, by using both an English corpus and new languages, we would be able to draw conclusions about the generalizibility of the proposed approach.

The English data is part of the Boston University radio news corpus [11], while the Catalan and Spanish recordings were taken from the news sub-part of the Glissando corpus [12]. The latter corpus already had syllable annotations, while for the English data this was derived from phone-level annotations, by applying English syllabification rules. A description of the characteristics of the three datasets is provided in Table 1.

Language	Type	Duration	Spkrs. (F+M)
Catalan	news	6hrs	8 (4+4)
English	news	3hrs	6 (3+3)
Spanish	news	6hrs	8 (4+4)

Table 1. *Description of the three datasets used in the experiments.*

3.2. Evaluation

The proposed system was evaluated both in terms of the goodness of the obtained syllable nuclei, as well as the placement of the syllable boundaries. The evaluation of the syllable nuclei is performed similarly to [7]: the middle of the frame having the highest nucleus sonority is considered as the position of the nucleus. If it falls within a reference vowel, it is considered correct, otherwise a deletion. If several nuclei are found inside a vowel, all but one are considered as insertions. The accuracy is then computed by subtracting the number of insertions from the correctly determined nuclei.

The syllable boundaries were evaluated in a similar manner. An automatic boundary was found to be correct if placed within 40 ms of a reference boundary, otherwise marked as substitution if found after the previous boundary/before the next boundary. All automatic markers found between two correct/substituted markers are considered as insertions. Boundaries having no corresponding automatic markers represent deletions. An evaluation example is illustrated in Figure 2. A measure of accuracy, similar to the one computed for nuclei detection, was then derived.

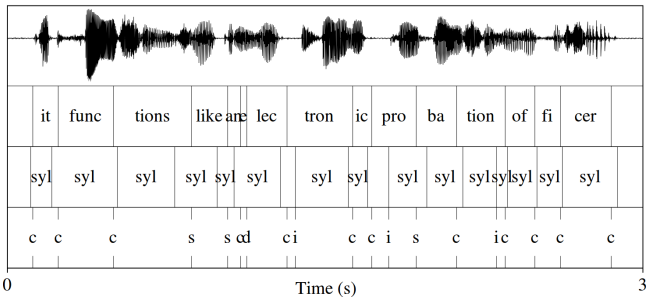

Fig. 2. Waveform and segmentation of the phrase "It functions like an electronic probation officer.". The upper tier contains the reference syllable segmentation, the middle tier the automatic segmentation, while the lower tier the corresponding evaluation (c=correct, s=substitution, d=deletion, i=insertion).

3.3. Results

The proposed system was compared against an open-source syllable segmentation tool, based on the energy of the signal [5], which will be further called *baseline*. The baseline system uses the energy function to detect syllable nuclei (peaks) and syllable boundaries (valleys), in conjunction with information about the harmonicity of the signal and its fundamental frequency.

The results obtained for nuclei detection with the proposed approach (using three different acoustic models) and the baseline are illustrated in Table 2. One can see that the proposed system outperforms the energy-based method for all

the languages and acoustic models used (except for *mono*, on both Catalan and Spanish).

Lang.	Syst.	Corr.	Del.	Ins.	Acc.
Catalan	base	.686	.314	.159	.527
	mono	.530	.470	.046	.484
	triA	.755	.245	.116	.639
	triB	.780	.220	.119	**.661**
English	base	.702	.298	.223	.479
	mono	.808	.192	.149	**.659**
	triA	.831	.169	.286	.545
	triB	.844	.156	.251	.593
Spanish	base	.680	.320	.166	.514
	mono	.444	.556	.052	.392
	triA	.727	.273	.099	.628
	triB	.769	.231	.103	**.666**

Table 2. *Nuclei detection results obtained on the three languages, for the baseline and the proposed approach using different acoustic models to obtain the broad phonetic class information.*

A similar picture can be observed when comparing the results of boundary placement (see Table 3). The syllable boundary performance is higher, compared to the baseline system, for the same acoustic models that outperformed the baseline for nuclei detection.

Lang.	Syst.	Corr.	Subst.	Del.	Ins.	Acc.
Catalan	base	.608	.198	.194	.067	.541
	mono	.427	.171	.401	.005	.422
	triA	.675	.158	.167	.050	.625
	triB	.708	.151	.141	.057	**.651**
English	base	.588	.269	.143	.093	.495
	mono	.628	.259	.113	.043	**.585**
	triA	.662	.296	.042	.137	.525
	triB	.669	.286	.045	.106	.563
Spanish	base	.666	.145	.189	.052	.614
	mono	.400	.123	.477	.002	.398
	triA	.660	.140	.201	.045	.615
	triB	.706	.135	.158	.043	**.663**

Table 3. *Syllable segmentation results obtained on the three languages, for the baseline and the proposed approach using different acoustic models to obtain the broad phonetic class information.*

4. DISCUSSION AND CONCLUSIONS

We have proposed a sonority-based method for syllable segmentation that outperforms an off-the-shelf energy-based system. We have found this to be true both for nuclei detection and boundary placement, on the language on which the recognizer was trained and also on two unseen languages (for most acoustic models tested). While the best results were not obtained with the same acoustic models (English seems to favour the *mono* model, due to its lower insertion rate), the *triB* model is the overall best, as it outperforms the baseline for each language and offers similar performance to *mono* on English. The results obtained are encouraging, the system proposed having a good generalizability. This characteristic would be especially useful for languages which do not have enough annotated resources to build a phonetic recognizer from which syllable segmentation can be derived.

An interesting observation can be made from the results in Tables 2 and 3: while the best model gives similar results for nuclei detection in the three languages, English has worse results than the other languages for boundary placement; this may be due to the existence of complex syllables in English (as in the word 'strengths'), which could make boundaries more difficult to locate. The World Atlas of Language Structures [13] classifies the syllabic structure of English as "complex" and Catalan and Spanish as "moderately complex" (although the former is considered to be more complex than the latter [14]). Among the 486 languages reviewed by the atlas for syllable structure 61 are considered as having a simple structure, 274 a moderately complex structure and 151 a complex structure. Since the vast majority of the reviewed languages from Africa, Asia and Latin America, where a high percentage of under-resourced languages are located, have at most a moderately complex syllable structure, we believe that out system could be used successfully in those languages.

The model that gave the best performance overall (*triB*) uses information about the identity of the speaker. While this can be an issue for under-resourced languages, there are several ways in which this can be overcome. For example, current speaker diarization systems can reach performances of almost 90% accuracy (see [15]), while at the same time we can use an utterance-based adaptation in Kaldi, thus eliminating the need for speaker identity.

From the results in Section 3.3, one can see that syllable boundary performance is highly correlated not only to the complexity of the syllabic structure of the language, but also to the quality of the obtained nuclei. The work presented in this study was a preliminary study on the usefulness of broad phonetic class information for syllable segmentation, so no particular optimizations were performed on the trained models. We will investigate in the future whether more optimized models or a finer sonority scale for vowels (low vowels>mid vowels>high vowels) would improve results.

One important issue that needs to be explored is how automatically detected syllables perform when used for other automatic tasks, like stress or prosodic boundary detection. Since the syllables found this way might not completely overlap phonologically defined syllables (due not only to the errors of the automatic syllable detection process, but also to syllabification rules that do not respect the sonority principle), it would be interesting to see the effect of these errors.

5. REFERENCES

[1] Hartmut Pfitzinger, "Local speech rate as a combination of syllable and phone rate," in *Proceedings of ICSLP*, 1998, paper 0523.

[2] Bogdan Ludusan, Antonio Origlia, and Francesco Cutugno, "On the Use of the Rhythmogram for Automatic Syllabic Prominence Detection," in *Proceedings of Interspeech*, 2011, pp. 2413–2416.

[3] Bogdan Ludusan and Emmanuel Dupoux, "Towards low-resource prosodic boundary detection," in *Proceedings of SLTU*, 2014, pp. 231–237.

[4] Paul Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[5] Antonio Origlia and Iolanda Alfano, "Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification.," in *Proceedings of LREC*, 2012, pp. 997–1002.

[6] George N. Clements, "The role of the sonority cycle in core syllabification," in *Papers in Laboratory Phonology*, John Kingston and Mary E. Beckman, Eds., pp. 283–333. Cambridge University Press, 1990.

[7] Jiahong Yuan and Mark Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *Proceedings of ICASSP*. IEEE, 2010, pp. 4222–4225.

[8] Jean-Philippe Goldman, "EasyAlign: An automatic phonetic alignment tool under Praat," in *Proceedings of Interspeech*, pp. 3233–3236.

[9] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of ICASSP*, pp. 5206–5210.

[11] Mari Ostendorf, Patti J. Price, and Stefanie Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[12] Juan María Garrido, David Escudero, Lourdes Aguilar, Valentín Cardeñoso, Emma Rodero, Carme De-La-Mota, César González, Carlos Vivaracho, Sílvia Rustullet, Olatz Larrea, et al., "Glissando: A corpus for multidisciplinary prosodic studies in Spanish and Catalan," *Language resources and evaluation*, vol. 47, no. 4, pp. 945–971, 2013.

[13] Martin Haspelmath and Matthew S. Dryer, "The world atlas of language structures online," 2008.

[14] Pilar Prieto, "The intonational phonology of Catalan," in *Prosodic typology II: The Phonology of Intonation and Phrasing*, Sun-Ah Jun, Ed., pp. 43–80. Oxford University Press, 2014.

[15] Sree Harsha Yella and Andreas Stolcke, "A comparison of neural network feature transforms for speaker diarization," in *Proceedings of Interspeech*, 2015, pp. 3026–3030.