SYLL-O-MATIC: AN ADAPTIVE TIME-FREQUENCY REPRESENTATION FOR THE AUTOMATIC SEGMENTATION OF SPEECH INTO SYLLABLES

Nicolas Obin, *François Lamare*, *Axel Roebel* †

IRCAM-CNRS UMR 9912-STMS
Paris, France
nobin@ircam.fr

ABSTRACT

This paper introduces novel paradigms for the segmentation of speech into syllables. The main idea of the proposed method is based on the use of a time-frequency representation of the speech signal, and the fusion of intensity and voicing measures through various frequency regions for the automatic selection of pertinent information for the segmentation. The time-frequency representation is used to exploit the speech characteristics depending on the frequency region. In this representation, intensity profiles are measured to provide information into various frequency regions, and voicing profiles are measured to determine the frequency regions that are pertinent for the segmentation. The proposed method outperforms conventional methods for the detection of syllable landmark and boundaries on the TIMIT database of American-English, and provides a promising paradigm for the segmentation of speech into syllables.

Index Terms : speech segmentation, syllable segmentation, time-frequency representation, information fusion.

1. INTRODUCTION

The segmentation of speech into segments is crucial in many applications of speech technologies (speech-to-text and text-to-speech systems). The main requirement of these technologies is the conversion of speech into a linguistic sequence that can be interpreted by humans, or by natural language processing (NLP) for further processing (human-computer interaction, spoken dialogue systems). Consequently, research has mostly focus on the study of phoneme or word recognition which has lead to the development of well-established speech recognition systems (HTK [1], SPHINX [2], HTS [3]). Speech segmentation systems are generally based on hidden Markov Models (HMM) in which acoustic and language models are determined with regard to a specific language. Consequently, the system requires to be adapted to the linguistic system of a language - i.e. the development of NLPs specific to the desired language - and cannot be used for under-resourced languages.

More recently, studies on the use of speech prosody have revealed the role of syllable segments - widely referred as the elementary segment of speech prosody - in speech recognition and synthesis systems [4, 5, 6, 7]. Moreover, the segmentation of speech into syllables may be extremely useful to improve content-based voice conversion systems (identity conversion, emotion transformation, or speech-to-sing systems) without requiring the use of NLPs [8, 9]. Contrary to the phoneme system which is specific to a language, the syllable is universally defined in terms of acoustic sonority : a syllable segment is fully determined by a maximum of sonority (the vowel nucleus) surrounded by local minimums of sonority (aggregation of consonants to the vowel nucleus). Accordingly, a universal syllable segmentation system may be used regardless to any specific language.

†This study was supported by the European FEDER project VOICE4GAMES.

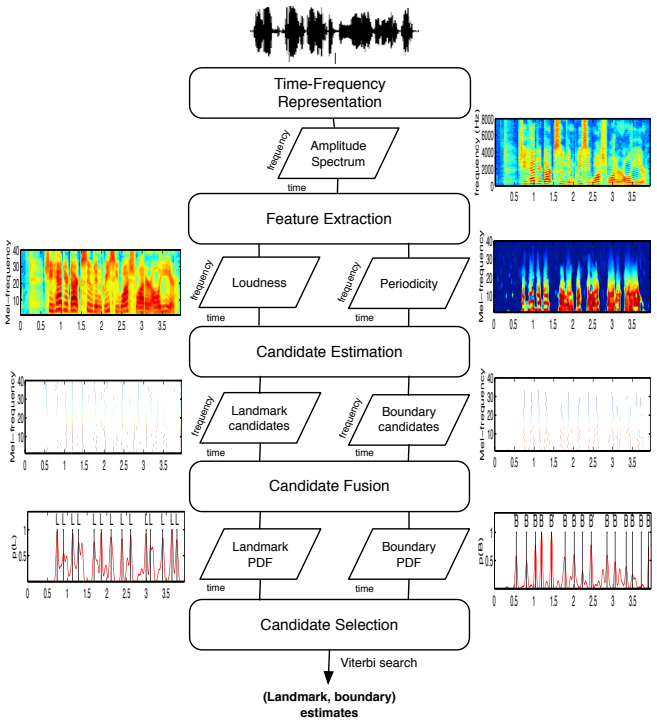

Fig. 1. Overview of the SYLL-O-MATIC system.

Most of existing syllable segmentation methods are derived from the MERMELSTEIN paradigms (from [10] to [11, 12]), optionally together with statistical processing (ANNs/HMMs [13, 14, 12]). In the MERMELSTEIN method [10], the measurement of sonority is approximated by the intensity measure through a spectral regions assumed to be relevant for the processing of vowel speech (formant region). Then, a recursive method is used for the final segmentation into syllables. More recently, the XIE and the ZHANG systems [12] have introduced the additional use of voicing information to improve the detection of vowel landmarks. Finally, the WU,

WANG, and the KALINLI systems [11, 15, 17] have investigated the exploitation of a multi-resolution spectral representation for syllable segmentation - using linear predictive coding, subband-based spectral correlation, and auditory attention cues. However, there is still no available well-established system for blind syllable segmentation.

This paper introduces novel paradigms for the blind segmentation of speech into syllables. The main idea of the proposed method is based on the use of a time-frequency representation of the speech signal, and the fusion of intensity and voicing measures through various frequency regions for the automatic selection of pertinent information for the segmentation. The time-frequency representation is used to exploit the speech characteristics depending on the frequency region. First, intensity and voicing profiles are determined over various frequency bands. Then, the voicing profile is used to determine the confidence that can be conferred to the corresponding intensity profile - i.e. for the selection of the spectral bands useful for the segmentation. Finally, intensity and voicing profiles are fused in order to determine an optimal profile that will be used for the segmentation.

2. WHAT IS A SYLLABLE ?

2.1. Definition

The syllable is a phonological unit of speech, which widely referred as the core element of speech prosody (speech rhythm and intonation). A syllable is typically composed of a nucleus (generally, a vowel) optionally surrounded by clusters of consonants (left and right margins) [18]. A syllable - pronounced "within a breath" -, is acoustically defined by the principle of sonority which is assumed to be maximal within the nucleus and minimal at the syllable boundaries. The definition of sonority is motivated by underlying physiological mechanisms (e.g., muscular tension, air flow, degree of co-articulation). which actually reflects the degree of organization and/or tension of speech.

Poo - poo - pee - doo !

Table 1. Typographical illustration of syllable segmentation, where hyphens indicate syllable boundaries.

2.2. Issues in Syllable Segmentation

In the idealistic - and extreme - case of hyper-articulated speech, each syllable would be clearly detached from each other by a silence. However, the confrontation with real-speech in real-world conditions causes a number of issues which introduced noisy information for the segmentation. In particular, articulation constraints introduce a number of noisy information for the identification of syllable landmark and boundaries. For instance, obstruent consonants (especially, occlusives) may introduce undesirable maximum in the intensity profile ; sonorant consonants may be confused with vowels ; partially voiced vowels (e.g., semi-vowels) may not be considered as a candidate for vowel landmark ; and the co-articulation of vowels across successive syllables may be extremely difficult to identify (e.g., CV + V). Finally, background noise and spontaneous speech introduces additional noise, and raise in articulation rate provide less-contrasted speech dynamics.

3. SYLL-O-MATIC

The objective of this study is to provide a robust measure of sonority based on the fusion of intensity and voicing measure in a single time-frequency representation. The issue of syllable segmentation is decomposed into landmark and boundary detection.

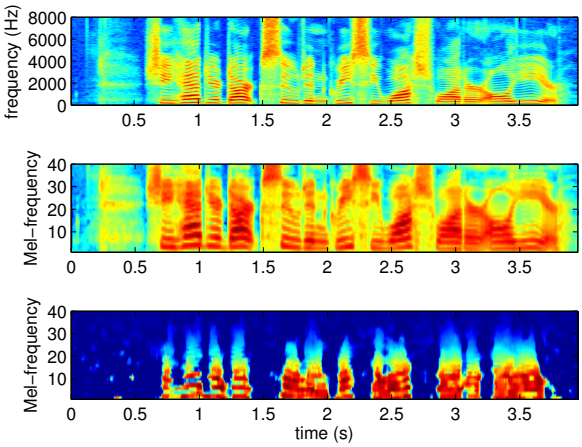

Fig. 2. Spectrogram, specific loudness, and voiced/unvoiced representations for the SA1 utterance : "*She had your dark suit in greasy wash water all year.*" spoken by FAKS0 speaker.

The segmentation is performed by exploiting a time-frequency representation of the speech signal. Landmark detection is performed by weighting the intensity information with the voicing information. Boundary detection is performed by exploiting the whole frequency information. Intensity and voicing profiles are measured over various frequency regions. Then, intensity and voicing profiles are fused in order to determine the final sonority profile used for the segmentation into syllables.

3.1. Multi-resolution Intensity Profiles

A time-frequency representation is used to measure the intensity contained into various frequency regions. For each frequency region, the specific loudness is measured as :

$$\mathrm{L}_t^{(k)} = \sum_{n=1}^{N^{(k)}} |A(t,n)|^{2\ 0.23} \tag{1}$$

where : k denotes the k-th frequency region, and $A(t,n)$ the amplitude of the n-th frequency bin at time t in the considered frequency region.

In this study, the specific loudness is measured over 40 Mel-frequency bands, with unitary integrated energy in order to enhance the information contained in low-frequency regions relatively to high-frequency regions. Then, the specific loudness $\mathrm{L}_t^{(k)}$ is normalized into a probability density function $\mathrm{L}_t^{(k)}{}_{\mathrm{norm}}$ so that each intensity profile will be further equally processed.

3.2. Multi-resolution Voicing Profiles

Also, a time-frequency representation is used in order to describe the degree of voicing into various frequency regions (VUV [19]). The analysis is based on a sinusoidal + noise representation of the signal [20]. Then, the degree of voicing of a particular frequency region is defined as the ratio of energy of the sinusoidal components observed in this region to the total energy of the frequency region [21, 22]. For each frequency region, the VUV is measured as :

$$\mathrm{VUV}_t^{(k)} = \frac{\sum_{j=1}^{N^{(j)}} |A_H(t,j)|^2}{\sum_{n=1}^{N^{(k)}} |A(t,n)|^2} \tag{2}$$

where : k denotes the k-th frequency region, $A_H(j)$ the amplitude of the j-th sinusoid, and $A(t,n)$ the amplitude of the n-th frequency

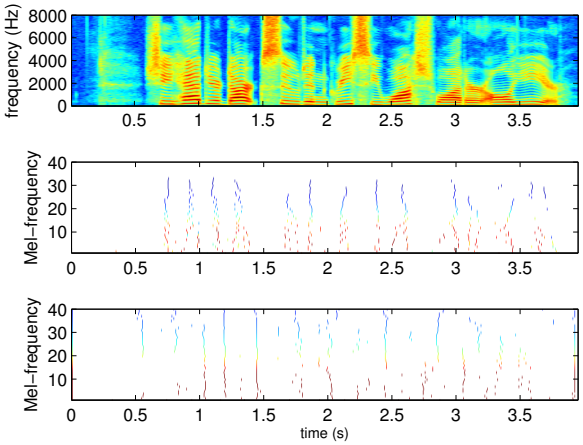

Fig. 3. Time-frequency positions of landmark (middle) and boundary (bottom) as determined for the SA1 utterance : "*She had your dark suit in greasy wash water all year.*" spoken by FAKS0 speaker.

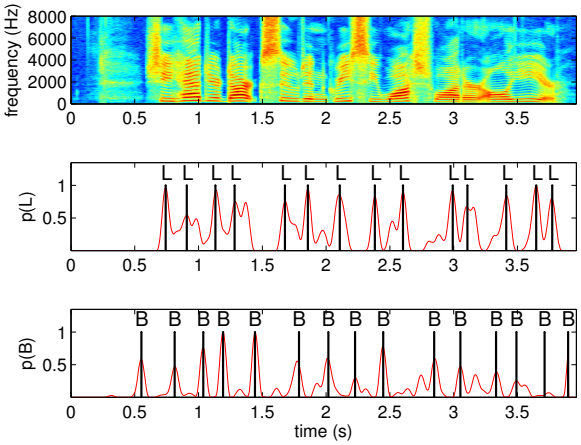

Fig. 4. Integrated landmark (middle) and boundary (bottom) time-position probabilities, and the determined sequence of landmarks (L) and boundaries (B) for the SA1 utterance : "*She had your dark suit in greasy wash water all year.*" spoken by FAKS0 speaker.

bin in the considered frequency band.

Hence, VUV is equal to zero when no harmonic content is present in the frequency band, and to one when only harmonic content is present in the frequency band. In this study, the VUV is determined over 40 Mel-frequency bands, with unitary integrated energy. An illustration of the time-frequency representation is provided in Figure 2.

3.3. Intermediate Fusion

Loudness and voicing profiles are then fused into a sonority profile so that the voicing profile is used as a confidence measure in the loudness profile observed at time t in the frequency region k :

$$\mathrm{S}_t^{(k)} = \mathrm{L}_t^{(k)}{}_{\mathrm{norm}} \times \mathrm{VUV}_t^{(k)} \tag{3}$$

This fusion is computed in order to select automatically useful information that will further be used for the detection of vowel landmarks - for which only voiced information is pertinent.

3.4. Candidates Selection

In the proposed time-frequency representation, the search for landmark positions exploits the degree of voicing as a mask to filter the spectral information, while the search for boundary positions exploits the whole spectral information. For each profile observed in a frequency region, time position of candidates for landmark and boundary are determined by using a simple method for minimum/maximum detection, respectively from the probabilities $\mathrm{S}_t^{(k)}$ and $\mathrm{L}_t^{(k)}{}_{\mathrm{norm}}$.

$$\mathrm{landmark}^{(k)} = \underset{t}{\mathrm{argmax}}\ \mathrm{S}_t^{(k)} \tag{4}$$

$$\mathrm{boundary}^{(k)} = \underset{t}{\mathrm{argmin}}\ \mathrm{L}_t^{(k)}{}_{\mathrm{norm}} \tag{5}$$

The candidate selection forms (KxT) matrices of landmark and boundary time/frequency positions (Fig. 3).

Then, the time positions of landmark and boundaries are determined through exploiting the time-frequency positions of candidates : the more frequent is observed a time position of a candidate over frequencies, the more likely is the presence of a landmark or a boundary. However, the exact time position of a marker may differ from one frequency region to the other - due to the asynchronism of the information contained in the frequency regions. Thus, landmark and boundary candidates are integrated over the frequency regions by using a moving average window (typically, a 20 ms. window), and then converted into a single probability density function p(L) and p(B) for landmark and boundary, respectively. Finally, the optimal sequence of landmark and boundary time positions is determined using a VITERBI search from the landmark and boundary probabilities (Fig. 4).

Optionally, a selection of relevant frequency bands is performed to regularize the integrated information for landmark detection.

$$\mathrm{p}_{\mathrm{norm}}(\mathrm{L}) = \frac{\mathrm{p}(\mathrm{L})}{\mathrm{N}_{\mathrm{voiced}}}, \qquad \max(\mathrm{p}_{\mathrm{norm}}(\mathrm{L})) = 1 \tag{6}$$

where : $\mathrm{N}_{\mathrm{voiced}}$ is the number of frequency regions explaining a certain amount of the total voicing of the analysis frame.

This is computed in order to re-enforce vowels observed within partially voiced regions - e.g., in the case of breathy/creaky syllables at the end of prosodic phrases (Fig. 5).

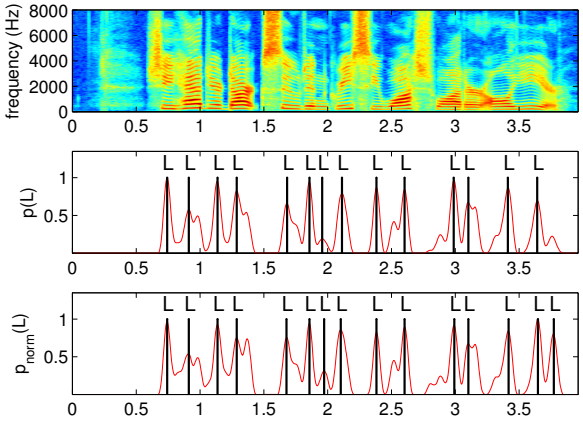

Fig. 5. Landmarks PDFs and the determined sequence of landmarks (L) with (bottom) and without (middle) the selection of useful frequency regions for the SA1 utterance : "*She had your dark suit in greasy wash water all year.*" spoken by FAKS0 speaker.

4. EVALUATION

The proposed method has been evaluated on the American-English TIMIT database. The TIMIT database is composed of 10 sentences read by 630 american-english speakers - which represents a total of 6300 speech utterances. The TIMIT database comes with phoneme and word alignment, but without syllable alignment. Thus, the reference syllable alignment has been obtained with the NIST syllabification software [23] - then, manually corrected. The TIMIT database is divided into a train (4620 utterances) and a test set (1680 utterances). For the present study and for comparison with supervised methods, the test set has been used for the evaluation. For exact comparison with [12], vowels and sonorants (/el/, /em/, /en/, /eng) were considered as a syllable nucleus. The average duration of a phoneme is around 80 ms. and the average duration of a syllable is 200 ms. The evaluation is decomposed into the detection of syllable landmarks (vowel nucleus) and the detection of syllable boundaries, with comparison to existing methods.

4.1. Landmark Detection

The landmark detection consists in the detection of the vowel region of the syllable. A landmark is considered as correct if it is detected within a syllable segment ([12]). The evaluation includes : MERMELSTEIN, XIE, fusion of landmark candidates based on specific loudness (MULTI-BAND), specific loudness and multi-band voiced/unvoiced measure (MULTI-BAND + VUV), and specific loudness, multi-band voiced/unvoiced measure and selection of relevant frequency bands for the fusion (MULTI-BAND + VUV + SELECTION). Additionally, the performance obtained from other methods - including statistical methods (ANNs/HMMs) - are reported from [12]. Finally, the comparison with the WANG and KALINLI methods are not reported due to large differences in the experimental setups. Insertion rate, deletion rate, and total error rate for the compared methods are reported in table 2.

TIMIT	INSERTION (%)	DELETION (%)	TER (%)
MERMELSTEIN	17.9	21.3	39.2
XIE	10.9	18.4	29.3
HOWITT	13.8	24.5	38.3
SPHINX 1	22.3	15.5	37.8
SPHINX 2	25.7	10.9	36.6
MULTI-BAND	19.8	10.1	29.9
MULTI-BAND + VUV	9.1	13.2	22.3
MULTI-BAND + VUV + SELECTION	**10.1**	**10.5**	**20.6**

Table 2. Performance for landmark detection on TIMIT.

4.2. Syllable Segmentation

The final objective of the segmentation of speech into syllables is to determine the time-positions of syllable onset and/or boundaries. The evaluation consisted in the comparison of the determined sequence of syllable boundaries to the reference one, with a +/- 50 ms tolerance (less than the average duration of a phoneme) on the exact position of the boundaries. The evaluation includes the MERMELSTEIN and XIE systems for a comparison with conventional methods. The implementation of the XIE system has been modified to determine the position of syllable boundaries : first, landmarks are detected based on periodicity and energy profiles as described in [12] ; then, syllable boundaries are determined by using the energy profile, only. Performances are reported in table 3.

TIMIT	INSERTION (%)	DELETION (%)	TER (%)
MERMELSTEIN	17.2	25.3	42.5
XIE	14.3	21.2	35.5
HOWITT	-	-	-
SPHINX 1	-	-	-
SPHINX 2	-	-	-
MULTI-BAND	24.7	10.1	34.8
MULTI-BAND + VUV	10.5	14.1	24.6
MULTI-BAND + VUV + SELECTION	**11.2**	**12.8**	**23.9**

Table 3. Performance for boundary detection on TIMIT.

4.3. Discussion

For the syllable segmentation, the detection of landmark is easier than the detection of boundaries. This observation is mostly due to the fact that the detection of boundaries is generally conditioned by the detection of landmarks, and that the consonant information is more noisy than the sonorant information. In particular, sonority peaks are generally more marked than sonority gaps - with exception of glides, dimly marked sonority peaks and partially voiced vowels.In all cases, the proposed method drastically outperforms all existing methods. The variants to the proposed method lead to the following conclusions : 1) the use of the whole frequency information presents a high rate of insertions which is due to noisy information contained in irrelevant frequency regions ; 2) the use of voicing information for the selection of relevant frequency regions significantly improves the segmentation ; 3) the selection of useful frequency regions successfully decreases the deletion of partially voiced vowel landmarks - with the counterpart of a slight increase of landmarks insertion.

5. CONCLUSION

In this paper, a time-frequency representation was introduced for the segmentation of speech into syllables. The main idea of the proposed method is based on the use of a time-frequency representation of the speech signal, and the fusion of intensity and voicing measures through various frequency regions for the automatic selection of pertinent information for the segmentation. The time-frequency representation is used to exploit the speech characteristics depending on the frequency region. In this representation, intensity profiles are measured to provide information into various frequency regions, and voicing profiles are measured to determine the frequency regions that are pertinent for the segmentation. The proposed time-frequency representation outperforms existing methods for the detection of syllable landmarks and boundaries, and provides a promising strategy for further research on the segmentation of speech into syllables.

6. REFERENCES

[1] S. J. Young, "The HTK Hidden Markov Model Toolkit : Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[2] K.-F. Lee, *Automatic Speech Recognition - The Development of the SPHINX System*. Kluwer, 1989.

[3] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based Speech Synthesis System Version 2.0," in *Speech Synthesis Workshop*, Bonn, Germany, 2007, pp. 294–299.

[4] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-Based Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358–366, 2001.

[5] T. Nagarajan, H. A. Murthy, , and R. M. Hegde, "Segmentation of Speech into Syllable-like Units," in *Eurospeech*, Geneva, Switzerland, 2003, pp. 2893–2896.

[6] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godjirey, "Modeling Prosodic Dynamics for Speaker Recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, 2003, pp. 788–791.

[7] N. Obin, "MeLos : Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, Ircam - UPMC, 2011.

[8] J. Tao, Y. Kang, and A. Li, "Prosody Conversion from Neutral Speech to Emotional Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, p. 1145–1154, 2006.

[9] E. E. Helander and J. Nurminen, "A Novel Method for Prosody Prediction in Voice Conversion," in *International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, 2007, pp. 509–512.

[10] P. Mermelstein, "Automatic Segmentation of Speech into Syllabic Units," *Journal of the Acoustic Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[11] S.-L. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating Syllable Boundary Information Into Speech Recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Munich, Germany, 1997, pp. 987–990.

[12] Z. Xie and P. Niyogi, "Robust Acoustic-Based Syllable Detection," in *International Conference on Spoken Language Processing*, Pittsburgh, USA, 2006, pp. 1571–1574.

[13] L. Shastri, S. Chang, and S. Greenberg, "Syllable Detection And Segmentation Using Temporal Flow Neural Networks," in *International Congress of Phonetic Sciences*, 1999, pp. 1721–1724.

[14] A. W. Howitt, "Automatic Syllable Detection for Vowel Landmarks," PhD. Thesis, Massachusetts Institute of Technology, 2000.

[15] D. Wang and S. Narayanan, "Robust Speech Rate Estimation for Spontaneous Speech," *IEEE Transactions on Audio, Speech, and Langage Processing*, vol. 15, no. 8, p. 2190–2201, 2007.

[16] Y. Zhang and J. R. Glass, "Speech Rhythm Guided Syllable Nuclei Detection," in *International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009, pp. 3797–3800.

[17] O. Kalinli, "Syllable Segmentation of Continuous Speech Using Auditory Attention Cues," in *Interspeech*, Florence, Italy, 2011, pp. 425–428.

[18] T. A. Hall, *Encyclopedia of Language and Linguistics*. Elsevier, 2006, vol. 12, ch. Syllable : Phonology.

[19] D. W. Griffin and J. S. Lim, "A New Model-Based Analysis/Synthesis System," in *International Conference on Acoustics, Speech, and Signal Processing*, Tampa, Florida, 1985, pp. 513–516.

[20] M. Zivanovic, A. Röbel, and X. Rodet, "Adaptive Threshold Determination for Spectral Peak Classification," *Computer Music Journal*, vol. 32, no. 2, pp. 57–67, 2008.

[21] N. Obin, "Cries and Whispers - Classification of Vocal Effort in Expressive Speech," in *Interspeech*, Portland, USA, 2012.

[22] N. Obin and M. Liuni, "On the Generalization of Shannon Entropy for Speech Recognition," in *IEEE workshop on Spoken Langage Technology*, Miami, USA, 2012.

[23] W. Fisher, "Tsylb Syllabification Package," 1996. [Online]. Available : ftp://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z