IDIAP RESEARCH REPORT

AN EMPIRICAL MODEL OF EMPHATIC WORD DETECTION

Milos Cernak Pierre-Edouard Honnet[a]

Idiap-RR-11-2015

JUNE 2015

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

An empirical model of emphatic word detection

Milos Cernak, Pierre-Edouard Honnet

Idiap Research Institute, Martigny, Switzerland

{milos.cernak,pierre-edouard.honnet}@idiap.ch

Abstract

The paper presents an empirical model of emphatic word detection, as an alternative to conventional machine-learning-based methods. The model is based on the Probabilistic Amplitude Demodulation (PAD) that is iteratively applied for getting syllable and stress modulations, i.e., using the cascaded PAD method. The emphatic words are detected by prominent peaks of the stress modulation and by considering the peaks that are stressed or accented. The cascaded demodulation steered with general purpose values derived from 200ms long average syllable duration, yields to detection accuracy of 81%–83%. Speaker-dependent cascaded demodulation, considering specific speaking rate of the speakers, yields to detection accuracy of 86%–91%. The advantages of the proposed empirical detection model are (i) noise-robustness, (ii) language-independence and (iii) it does not require a training phase.

Index Terms: speech emphasis, probabilistic amplitude demodulation

1. Introduction

In speech communication, we can change the stress, for example, from the principal noun to another content word, to call attention to what we want to emphasise. The changes observed at emphatic words can be exemplified by differences of prosodic features and in spectral domain as well. The detection and prediction of the emphatic words is important because the words we stress can change the underlying meaning. It is desirable for the applications such as speech-to-speech translation, to transfer also correct accent and stress prosody features, along with the word emphasis.

Previous attempts on emphatic word detection were primarily focused on modelling differences in fundamental frequencies [1] and overall intensity [2]. Later, the duration, spectral features [3], lexical features [4, 5], and word identity features [6] were proposed. We can find several studies in the literature published in the last two decades, with state-of-the-art detection accuracy for different languages of about 80%-90%. A conventional method is to extract acoustic and linguistic features, and apply machine learning to train a classifier (model).

We are interested in empirical modelling of emphatic word detection, that is (i) robust to noise, (ii) does not require training, and (iii) is language-independent. The empirical model of Tamburini [7] is an example, however the objective function of the prominence identification was ad-hoc and language dependent. On the other hand, the method was correctly based on the two different aspects of prosodic typology: the prominence and the rhythmical patterns of an utterance [8]. We hypothesise that also realisation of the emphatic word is achieved using the relative prominence of adjacent elements, and the rhythmical patterns. The question arises, how to get a reliable estimate of those prosodic features?

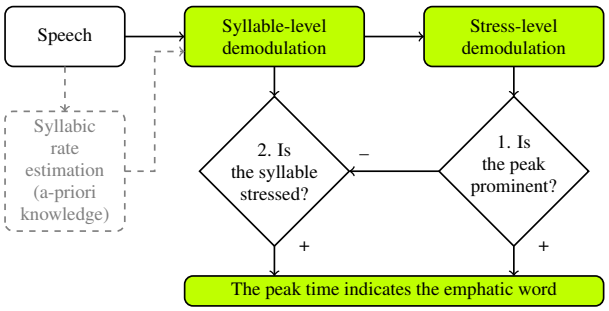

Figure 1: The empirical model of emphatic word detection. The emphatic word is detected using the most prominent stress modulation peak or the peak coming from the stressed syllable.

A speech signal conveys information on different time-scales. Traditionally, sequential speech processing suggests the segmental and supra-segmental time-scales to be used for different models of our interest, such as for the acoustic and prosodic modelling. Different time-scales have often been treated independently in the past. However, we can hypothesise that they are related, and that this relation is important also for the emphatic word detection.

To provide us an insight into the prosody hierarchy, we have selected the Probabilistic Amplitude Demodulation (PAD) approach [9]. The PAD method is noise robust and allows the algorithm to be steered using a-priori knowledge of modulation time-scales, i.e., the user can specify the prosodic tiers — stress, syllables, and utterance — to be analysed. And as an analytic model, it is assumed to be language independent. The PAD method can be used iteratively to get progressively slower prosodic tiers, as depicted in Fig. 1 that sketches the empirical observation of emphatic word detection. It was shown that rhythmical patterns can be reliably detected by phase relations of the syllable and stress modulations. The work of Leong et al. [10] argues that phase relations are much more important to the stress detection than the modulation amplitudes. We hypothesise that the second important aspect of the emphatic word detection – prominence – is related to the stress modulation amplitude, as it is related to the energy. The energy (or intensity) was already proved to be a very useful acoustic parameter [2].

We experimentally evaluate the proposed method within the SIWIS project – Spoken Interaction with Interpretation in Switzerland[1] on French and English speech data. The structure of the paper is as follows: the PAD method is introduced in Section 2. Section 3 describes the experimental setup and used speech database. The results are shown in Section 4. Finally the conclusions follow in Section 5.

[1] http://www.idiap.ch/project/siwis

2. Probabilistic amplitude demodulation

The Probabilistic Amplitude Demodulation (PAD) models the speech signal y_t as:

$$y_t = c_t.m_t \tag{1}$$

where c_t and m_t are a carrier and modulator components, respectively. The modulator is represented as a non-linear function

$$m_t = m(x_t) = \sigma_m \log(1 + \exp(x_t)) \tag{2}$$

of the transformed-modulator signal x_t, with the amplitude σ_m, drawn from a stationary Gaussian process. The covariance function of x_t represents the typical time-scale of variations of m_t, and importantly, it can be controlled manually using a-priori user knowledge. The carrier is modelled as a Gaussian process which is uncorrelated in time.

There are many solutions for solving Eq. (1). The PAD method describes a Bayesian inference given the data for extracting the amplitude modulation structure. More specifically, posterior probability of all the possible modulators and carriers given the data is:

$$p(c_1^T, m_1^T | y_1^T, \theta) = \frac{p(y_1^T, c_1^T, m_1^T | \theta)}{p(y_1^T | \theta)}, \tag{3}$$

where $p(y_1^T, c_1^T, m_1^T | \theta)$ is the joint probability of the signal, carrier and modulator, T is the number of frames of the processed speech signal, and θ corresponds to the model parameters. The most probable modulator and carrier are obtained by the *maximum a posteriory* (MAP) inference as:

$$\hat{c}_1^T, \hat{m}_1^T = \underset{c_1^T, m_1^T}{\operatorname{argmax}} \; p(c_1^T, m_1^T | y_1^T, \theta), \tag{4}$$

using a gradient-based method that is used to search for the optimal solution. To allow the demodulation to be user steerable, i.e., perform the demodulation using a specific time-scale, the parameters of the model θ can be obtained by the MAP inference as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; p(\theta | y_1^T) = \underset{\theta}{\operatorname{argmax}} \; p(y_1^T | \theta) p(\theta), \tag{5}$$

where the prior over parameters $p(\theta)$ is set by the user. The maximum-likelihood estimate is recovered when the prior is uniform, i.e., $p(\theta) = constant$.

To reveal different time-scale information present in the speech signal, we used the PAD process to decompose the signal into a cascade of modulators and a carrier [11]. The time-scales of the modulators are considered as the prior constants $p(\theta)$, creating a concept of steered demodulation. A first demodulation is performed with a syllable-based modulation where an average syllable duration in samples is used as the parameter prior $p_{syll}(\theta)$. The obtained syllable envelope σ_{syll} is used as input signal for progressively slower demodulation at the stress level, using a different prior $p_{stress}(\theta)$, to generate a stress envelope σ_{stress}. The general purpose values for the speech signal demodulation could be 5Hz for the first decomposition with the syllable frequency, and an average between the half and one third of the syllable frequency for the stress modulation frequency. For example, considering 16kHz sampled data, the values could be $p_{syll}(\theta) = 3200$ samples and $p_{stress}(\theta) = 8000$ samples. The better prior estimate of the syllabic rate, the more accurate the obtained cascaded demodulation.

In addition, the PAD method is able to deal with noisy data, as it explicitly incorporates additive uncorrelated Gaussian noise around the product of $c_t.m_t$.

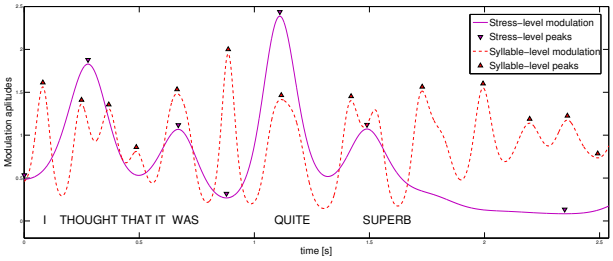

Figure 2: *An example of the PAD of an English sentence with the transcribed text on the bottom. Prosodic stress is detected by speech instances where the modulation signals are on phase (time-aligned peaks).*

2.1. Lexical stress detection

Recent work of Leong et al. [10] showed that stressed syllables can be reliably detected by phase relations of the σ_{syll} and σ_{stress} envelopes. Fig. 2 shows an example of the demodulated signals and their phase relation. The strong syllables (with lexical stress) are represented by modulation signals on phase – with the aligned peaks, and the weak syllables (unstressed) are represented by modulation signals with the misaligned peaks.

The accuracy of the lexical stress detection depends on correct time-scale specification. Using speaker-dependent values (inferred from the speech samples of the speakers) performs usually better than speaker-independent values (such as general purpose values based on an average syllable duration).

2.2. Emphatic word detection

Our proposal of the empirical model of emphatic word detection is shown in Fig. 1. The cascaded PAD is steered either by the speaker-independent time-scales of the modulators (i.e., the unknown priors) , or by the speaker-dependent time-scales (i.e., the known priors). The speaker-independent estimates may be adjusted for a specific language. Our test set consisted of French and English data. Although French syllabic rate is slightly higher than English syllabic rate, we use the same unknown prior for the syllable demodulation for both languages. Because the syllabic rate differs across languages [13] (for example Japanese has even faster syllabic rate than French), further adjustment is recommended.

The detection itself is straightforward. After the steered cascaded demodulation, the local maximum amplitude peaks are found and processed as follows:

1. Stress-level modulation amplitude σ_{stress}: If the global maximum is a prominent (relatively higher at least 15%) with respect to others – the time of the maximal peak indicates the emphatic word. The emphasised word "quite" of Fig. 2 is an example.
2. Syllable-level modulation amplitude σ_{syll}: If not, consider a group of the most prominent stress peaks, and select the peak that is synchronised with the σ_{stress}, i.e., the peak comes from the stressed or accented syllable (see Fig. 3).

The hypothesis is that emphasised speech is exemplified by the prominent syllable-based energy and contains a strongly stressed or accented syllable. Thus, we propose to detect both aspects of the prosodic typology: prominence related to the stress modulations amplitude, and rhythmical patterns related to the strong syllables.

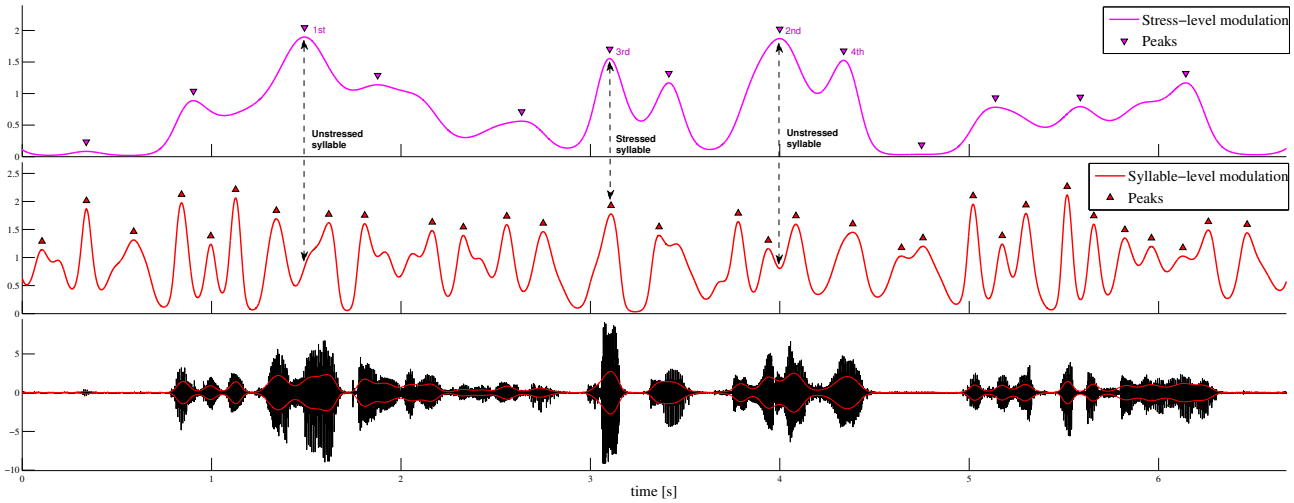

Figure 3: *An example of correct emphasised word detection (in capitals) of the French sentence "L'agriculture marocaine bénéficie d'un TRAITEMENT PRIVILÉGIÉ pour ses exportations vers l'Europe". From the first 4 most prominent peaks, the 3rd and 4th indicate the emphasised words.*

3. Experiments

The data used for these experiments is part of the evolving SIWIS database, from the SIWIS project [14]. It consists of 84 bilingual speakers. Each speaker was asked to read 180 prompts in two languages. The primary intended use of this bilingual parallel corpus is speech to speech translation. The recordings were done in a booth at the University of Geneva. About 25 prompts of the sentences – among which 5 questions – taken from Europarl corpus [15], were read twice: once with no specific instructions and another time with focus on one predefined word. The corresponding transcription for each sentence was given, with a tag on the words that the speakers were asked to emphasise.

Three speakers were selected for our experiments: one English male speaker *EN-B1_19*, one French male *FR-A1_19* and one French female *FR-A1_08*. We only used the sentences which contained the emphasis. We generated contextual labels using the French text analyser eLite [16] and an English text analyser of the Festival[2] system [17]. The contextual labels were aligned using forced alignment done by HTS [18]. The Viterbi algorithm is used to estimate phone boundaries of the speech to be aligned. For French, the aligning models were trained using speaker adaptive training [19] on the BREF database [20] using a total of about 13000 sentences, coming from 100 speakers. For English, the models were trained on about 7000 sentences of the Wall Street Journal database [21], coming from 166 speakers.

An additional contextual feature was given to make the distinction between emphasised and non emphasised words. These time aligned labels were then used to assess the validity of our approach in the task of detecting which word had the main emphasis in each sentence. The test data amounted to 65 sentences corresponding to 942 syllables:

- 42 sentences for French (21 per speaker), representing 584 syllables.
- 23 sentences for English corresponding to 358 syllables.

[2]http://www.festvox.org/festival/

Finally, the test was performed using the proposed emphatic word detection method, described in Sec. 2.2, using:

1. Unknown time-scales of the modulators, the priors $p_{syll}(\theta)$ and $p_{stress}(\theta)$, where we used general purpose values derived from 5Hz syllable frequency (cf. Sec. 2).
2. Known time-scales of the modulators derived from the forced aligned labels: syllable frequencies for *EN-B1_19* – 4.4 Hz, *FR-A1_19* – 4.8 Hz and *FR-A1_08* – 4.6 Hz.

4. Results

Tab. 1 shows detection accuracy for both cases. The results indicate that (i) the steering the PAD using the known time-scales (speaker-dependent priors) improves detection, and (ii) detection on English data performs better. However, a t-test shows that the differences between the results are not statistically significant ($p > 0.05$).

Table 1: *Accuracy of emphatic word detection.*

Test	Unknown time-scale	Known time-scale
French	80.95%	**85.71%**
English	82.61%	**91.30%**

Improvement using the known priors was expected. Therefore the authors of the PAD method proposed learning of the parameters in their work [9]. In their study [10], Leong et al. estimated the priors using the modulation filter-bank comprising of "Stress" and "Syllable" finite-impulse response band-pass filters, 0.8–2.3 Hz and 2.3–7 Hz, respectively. We estimated the average syllabic rate from the forced aligned labels. The stress time-scale was then estimated as:

$$p_{stress}(\theta) = \frac{2.p_{syll}(\theta) + 3.p_{syll}(\theta)}{2}. \quad (6)$$

The difference between languages is related to the prominence of the stress modulation amplitude σ_{stress}. For English, 75% of correct detection cases were attributed to the prominent

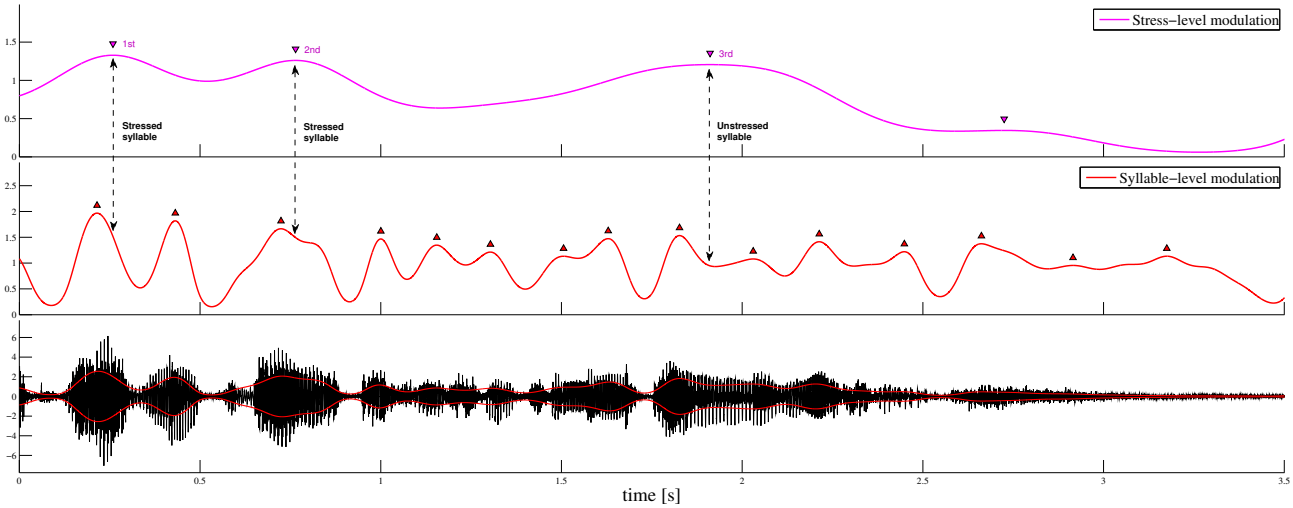

Figure 4: *An example of incorrect emphasised word detection (in capitals) of the English sentence "What we are doing today is essentially a NUISANCE". From the first 3 most prominent peaks, the 2nd was wrongly detected as the emphasised word (the 3rd was correct).*

σ_{stress}, and 25% to the rhythmical patterns (i.e., the prominent peak coming from the stressed syllable). For French, the σ_{stress} was significant in 56% cases and the rhythmical patterns increased to 44% cases.

Fig. 3 shows an example of correct emphasised word detection of a French sentence. Both emphasised words are detected using the rhythmical patterns. Fig. 4 shows an example of incorrect emphasised word detection of an English sentence. We see that the width of the prominent stress modulation peaks may also be important, which could be related to the longer duration of the emphatic words.

5. Conclusions and future directions

We have proposed an empirical model of emphatic word detection that

1. does not require training,
2. is robust to noise because of the PAD method,
3. and as a bottom-up method, is taken to be language-independent.

The detection model gives state-of-the-art performance using the unknown time-scale priors, i.e., suitable for speaker-independent tasks. The use of the PAD method was motivated by the hypothesis that the emphatic words can be detected by examining the prosody hierarchy; more specifically, by relation of different time-scale information conveyed in speech. We clearly demonstrated that the relation of the stress and syllable time-scale information is important for the emphatic word detection. In addition to their synchronisation, reported earlier to be important for stressed syllable detection [10], the stress modulation amplitude has the most significant impact on the emphatic word detection.

A third parameter we observed, the width of the stress modulation peaks, can be considered as an another acoustic correlate to duration of the emphatic words. It may be noted that this issue can be related to the differences between stress-timed and syllable-timed languages. While in stress-timed languages, such as English, we give stress to certain words while others are spoken more quickly, in syllable-timed languages, such as French, syllables receive equal importance. Therefore we may speculate that the width of the stress modulation peaks would work for English, such as for correct detection of 3rd prominent peak that belongs to an emphatic word in the example of Fig. 4. However, the peak width may not work well for French where emphasised words do not have distinctive duration. Further analysis is required to evaluate an impact of the stressed/accented peaks and widths of the peaks (of the syllable-level modulation) on the emphatic word detection. We have not investigated this in the current study due to lack of testing data.

There are different types of emphasised words:

1. intonation accent where the word is prominent within a prosodic phrase,
2. emphatic stress investigated in this paper,
3. constrastive stress used to point out differences between words,
4. and new information stress when asked a question, the requested information is naturally emphasised.

These stress types differ in their linguistic interpretation, however we believe that all the types share their acoustic properties. Therefore we consider our proposed method to be suitable for any stress type.

We plan to extend the testing to more speakers and languages, with intention to use it for the prosody transfer technology we work on, similarly to the work of Anumanchipalli et al. [22], but replacing the cross-lingual accent analysis by the proposed empirical emphatic word detection. We further plan to extend the model with the width analysis of the prominent peaks of the stress modulation amplitude.

6. Acknowledgements

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS), and under SP2: the SCOPES Project on Speech Prosody.

7. References

[1] D. R. Ladd and R. Morton, "The perception of intonation emphasis: Continuous or categorical?" *Journal of Phonetics*, vol. 25, pp. 313–342, 1997.

[2] M. Heldner, E. Strangert, and T. Deschamps, "Focus Detection Using Overall Intensity and High Frequency Emphasis," in *Proc. of ICPhS*, 1999.

[3] M. Heldner, "Spectral emphasis as an additional source of information in accent detection," in *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, M. Bacchiani, J. Hirschberg, D. Litman, and M. Ostendorf, Eds. ISCA, 2001, pp. 57–60. [Online]. Available: www.speech.kth.se/prod/publications/files/710.pdf

[4] J. M. Brenier, D. M. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Proc. of Eurospeech*, 2005, pp. 3297–3300.

[5] A. Nenkova and D. Jurafsky, "Automatic detection of contrastive elements in spontaneous speech," in *Proc. of ASRU*. IEEE, Dec. 2007, pp. 201–206. [Online]. Available: http://dx.doi.org/10.1109/asru.2007.4430109

[6] A. Margolis and M. Ostendorf, "Acoustic-based pitch-accent detection in speech: Dependence on word identity and insensitivity to variations inword usage," in *Proc. of ICASSP*, vol. 0. Los Alamitos, CA, USA: IEEE, Apr. 2009, pp. 4513–4516. [Online]. Available: http://dx.doi.org/10.1109/icassp.2009.4960633

[7] F. Tamburini, "Automatic Prominence Identification and Prosodic Typology," in *Proc. of Interspeech*, 2005, pp. 1813–1816.

[8] S.-A. Jun, "Prosodic Typology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford University Press, 2005, pp. 430–458.

[9] R. E. Turner and M. Sahani, "Demodulation as Probabilistic Inference," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2398–2411, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1109/tasl.2011.2135852

[10] V. Leong, M. A. Stone, R. E. Turner, and U. Goswami, "A role for amplitude modulation phase relationships in speech rhythm perception." *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 366–381, Jul. 2014. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/24993221

[11] R. E. Turner and M. Sahani, "Probabilistic amplitude demodulation," in *Proc. of Independent Component Analysis and Signal Separation*, 2007, pp. 544–551. [Online]. Available: http://www.gatsby.ucl.ac.uk/ turner/Publications/TS2007ICA.pdf

[12] A. Malécot, R. Johnston, and P. A. Kizziar, "Syllabic rate and utterance length in French." *Phonetica*, vol. 26, no. 4, pp. 235–251, 1972. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/4670762

[13] F. Pellegrino, C. Coup, and E. Marsico, "A cross-language perspective on speech information rate," *Language*, vol. 87, no. 3, pp. 539–558, 2011.

[14] P. N. Garner, R. Clark, J.-P. Goldman, P.-E. Honnet, M. Ivanova, A. Lazaridis, H. Liang, B. Pfister, M. S. Ribeiro, E. Wehrli, and J. Yamagishi, "Translation and prosody in Swiss languages," in *Nouveaux cahiers de linguistique française*, 2014.

[15] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. of 10th Machine Translation summit*, 2005, pp. 79–86.

[16] S. Roekhaut, S. Brognaux, R. Beaufort, and T. Dutoit, "eLite-HTS: A NLP tool for French HMM-based speech synthesis," in *Proc. of Interspeech*, 2014, pp. 2136–2137.

[17] A. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," Human Communication Research Centre, University of Edinburgh, Technical Report, 1997.

[18] 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 *Proc. of ISCA SSW6*, 2007, pp. 131–136.

[19] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1109/tasl.2008.2006647

[20] L. F. Lamel, J.-L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for French," in *Proc. of Eurospeech*, 1991, pp. 505–508.

[21] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, ser. HLT '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 357–362. [Online]. Available: http://dx.doi.org/10.3115/1075527.1075614

[22] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Intent transfer in speech-to-speech machine translation," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, Dec. 2012, pp. 153–158. [Online]. Available: http://dx.doi.org/10.1109/slt.2012.6424214