Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

The Impact of Word Alignment on Audio-visual Word Prominence Detection

Martin Ernst Heckmann, Paschalis Mikias, Dorothea Kolossa

2014

Preprint:

This is an accepted article published in 11. ITG Conference on Speech Communication. The final authenticated version is available online at: https://doi.org/[DOI not available]

The Impact of Word Alignment Accuracy on Audio-visual Word Prominence Detection

Martin Heckmann, Paschalis Mikias, Dorothea Kolossa

Honda Research Institute Europe GmbH, D-63073 Offenbach Main, Germany
Email: {martin.heckmann}@honda-ri.de
Web: www.honda-ri.de
Cognitive Signal Processing Group, Ruhr-Universität Bochum, D-44780 Bochum, Germany
Email: {paschalis.mikias, dorothea.kolossa}@rub.de
Web: www.ruhr-uni-bochum.de

Abstract

To automatically detect prominent syllables or words, most approaches require a segmentation of the speech signal and a subsequent extraction of prosodic features in these segments. In this paper we investigate the impact of the precision of this segmentation on the detection. We perform the segmentation of our audio-visual prosodically rich corpus based on an HMM trained on a large dataset. Thereby, we investigate different training strategies of the HMM. We consider on one hand training without any prior information, i.e. flat start and on the other hand when using partially manually created segmentations. Additionally we also introduce features tailored to detect onsets in the spectrogram. We evaluate the performance of the segmentation on our corpus on one hand by comparing it to manual annotations and on the other hand functionally, i. e. via the impact on the prominent word detection. The results show that the use of manual annotations in the training and the onset features significantly improve the segmentation accuracy. Yet the results of the prominent word detection do not to benefit from the better segmentation. From this we conclude that the extraction of the prosodic features is robust against segmentation errors.

Index Terms: segmentation, Gabor filter, audio-visual, word prominence, contour, FPCA

1 Introduction

Prosodic variations in the speech signal are very important in human communication. For example after a misunderstanding humans make the corrected word very prominent [1]. Despite its low penetration in commercial products, the prosocid analysis of the speech signal is also a very important research topic [2]. Enabling a machine to exploit this information is the target of our and previous research [1, 3]. The acoustic correlates of prominence are usually a longer duration as well as specific pitch and intensity patterns [4, 5]. To detect these prominent words, in most cases functionals, e. g. mean or max, are extracted from the acoustic features [1, 3, 6]. The calculation of these functionals requires a segmentation of the words or syllables. For this task only relatively few approaches have been proposed [7–9]. The most common approach is to perform the segmentation with an HMM-based speech recognition system.

In this paper we investigate two different HMM training strategies and a novel feature set tailored to detect onsets in the spectrogram. We evaluate the resulting segmentations on one hand with manual annoatations and on the other hand functionally, i. e. via the performance of the word prominence detection we obtain with the different segmentations. For the evaluation we use an audio-visual corpus containing user corrections. Via the inclusion of visual information we in particular hope to obtain more robust detections. Information on the movements of the speaker's mouth and face notably improves the accuracies of automatic speech recognition, particularly in difficult situations [10–13]. On the other hand humans are also able to use such visual information to extract prosodic cues [14–16].

In the next section we introduce the dataset we used for the evaluation. We will detail the different word HMM training methods and our onset features in Section 3. After that Section 4 describes the different features extracted from the acoustic and visual channel. Following this Section 4 will present the results of the evaluation. Then we will discuss the results in Section 5 and give a conclusion in Section 6.

2 CoCo Corpus

Our target scenario is the detection of corrected words in human-machine interaction via prosodic cues. To simulate this we recorded subjects interacting via speech in a Wizard of Oz experiment with a computer in a small game where they moved tiles to uncover a cartoon [17]. This game yielded utterances of the form 'place green in B one'. Occasionally, a misunderstanding of one word of the sequence was triggered and the corresponding word highlighted, verbally and visually. The subjects were told to repeat in these cases the phrase as they would do with a human, i. e. emphasizing the previously misunderstood word. However, they were not allowed to deviate from the sentence grammar by e. g. beginning with 'No'. This was expected to create a narrow focus condition (in contrast to the broad focus condition of the original utterance) thereby making the corrected word highly prominent. We will refer to this corpus as the *Correction Corpus (CoCo)*. Following a manual annotation by 3 human annotators, from the originally 16 speakers, we retained only those where the inter-annotator agreement measured by Fleiss' kappa was at least moderate, i. e. $\kappa > 0.5$ [18]. This yields 12 speakers, 6 females and 6 males. The audio signal was originally sampled at 48 kHz and later downsampled to 16 kHz. For the video images a CCD camera with a resolution of 1280×1024 pixels and a frame rate of 25 Hz was used.

For further processing those turns where the original utterance and a correction were available were selected. Overall we have 2023 turn pairs (original utterance + correction), i. e. on average ≈ 160 turn pairs per speaker. From these the word which was emphasized in the correction was determined. Then it was extracted as well in the original utterance as in the correction. This yields a dataset with each individual word taken from a broad and a narrow focus condition.

3 Word Alignment

To determine the word boundaries we performed a forced alignment with an HMM. In particular we decoded the utterance while providing the word sequence during decoding. Hence the only parameter to be determined by the system is the temporal alignment. We used HTK [19] and trained it on the Grid Corpus [20]. The Grid Corpus contains 34000 utterances from 34 speakers which we equally divided for training and testing. For the first training approach which we investigated we assumed that apart from the text transcription no prior information is a available. Hence we initialized the HMM models uniformly and the mapping of HMM states to observed features is then iteratively refined during training via the Baum-Welch algorithm. This is also termed *Flat Start*. In the second approach we used the temporal alignments provided by the Grid corpus for the initialization of the HMM states. We will refer to this initialization as *Segmented*. For both approaches we used 39-dimensional MFCCs as feature

ISBN 978-3-8007-3640-9

© VDE VERLAG GMBH · Berlin · Offenbach

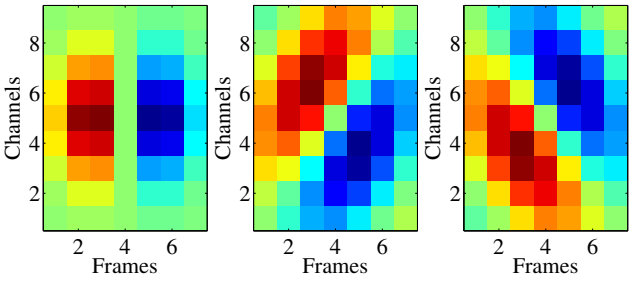


Figure 1: Gabor filters used to detect temporal onsets. See text for more details.

vectors [21]. They contain 13 cepstral coefficients and their first and second temporal derivative. We also investigated the impact of a speaker adaptation with a Maximum Likelihood Linear Regression (MLLR) step with a subsequent Maximum A-Posteriori (MAP) step on the alignment. For the adaptation we used all the data available for a given speaker, i. e. not partitioning in training and test set.

3.1 Onset Features

MFCCs are designed to best capture the phonetic information of the speech signal. This does not necessarily mean that they are also best suited for temporal segmentation. We therefore also investigated a different feature set. For these features we combined three different approaches. Inspired by an approach for syllable segmentation [7] we calculated the short term energy of the speech signal and determined the position of the minima in the low-pass filtered signal. The feature we derived from that is 1 at the position of a minimum and 0 at all other instances. To avoid numerical instabilities during the training of the HMM we added Gaussian noise with zero mean and a variance of .005 to the feature. The second approach is the calculation of a temporal derivative in the spectrogram. From this we expected to capture the onsets of syllables and words. As in [22] we applied a 100-channel Gammatone filterbank to the signal, calculated the envelope and performed a smoothing along the frequency axis with a Mexican Hat filter. We then filtered this pre-processed spectrogram with a 2D Gabor filter [22] using only the imaginary part of three Gabor filters with different orientations (compare Fig. 1). We used orientations of $90°, 45°$ and $135°$, a width of 70 ms (7 frames) and a height of 9 Gammatone channels for the Gabor filters as we expect these filters to well capture onsets as well as formant transitions, in particular occurring at word boundaries. This filtering yielded three spectrograms. At each time-frequency unit we retained only the response of the strongest filter. Hence we again obtained one spectrogram. As final feature, we then retained the sum of all of these values along the frequency axis for each instant in time. To obtain our final feature set we combined the minima and the Gabor feature with the MFCCs. This yielded a 41-dimensional feature vector. We want to refer to these features as Onset Features.

4 Prosodic Features

We extracted different features from the acoustic and visual channel to capture the prosodic variations [18].

4.1 Acoustic features

Since we expected the loudness l to better capture the perceptual correlates of prominence than the energy we extracted it by filtering the signal with an 11th order IIR filter as described in [23], followed by the calculation of the instantaneous energy, smoothing with a low-pass filter with a cut-off frequency of 10 Hz, and conversion into dB. Furthermore, we calculated D, the duration of the word and the gaps before and after the word as determined from the forced alignment. We also extracted the fundamental frequency f_0 (following [24]), interpolated values in the unvoiced regions via cubic splines and converted the results to semitones. To detect voicing we used an extension of the algorithm described in [25]. Finally, we also determined the spectral emphasis SE, i. e. the difference between the overall signal energy and the energy in a dynamically low-pass-filtered signal with a cut-off frequency of $1.5 f_0$ [26].

4.2 Visual features

To extract features from the visual channel we used the OpenCV library [27] to detect the nose in each image. Based on this we developed a tracking algorithm which yields the nose position over time. The nose does not move much during articulation relative to the head and is hence well suited to measure the rigid head movements. Starting from the nose we can determine the mouth region in the image via a fixed speaker independent offset. On each subsampled mouth image of size 100×100 pixels we calculated a two-dimensional Discrete Cosine Transform (DCT). Out of the 10000 coefficients per image we selected the 20 with the lowest spatial frequencies.

4.3 Functionals

As post-processing step we normalized the prosodic features by their utterance mean and calculated their first and second derivative (except for duration). We then extracted as functionals the mean, max, min, spread (max-min) and variance along the word. Word boundaries were determined by the alignment described in Sec. 3.

4.4 FPCA Contour Models

Most approaches in the computational processing of prosody rely on functionals derived from acoustic features [6, 28]. On the other hand the contours of the features also contain important information. We found that Functional Principal Component Analysis (FPCA) is a very well suited contour model for the task of word prominence detection [18]. Functional data analysis is a branch of statistics which operates on curves or functions instead to data points [29–31]. Prior to applying FPCA a given contour $\mathbf{s}$ has to be interpolated to a new contour defined on an equally spaced grid of fixed length and represented as a row vector $\mathbf{x} = [x(1), \ldots, x(N)]$. The first step in FPCA is then the smoothing of the data. This transforms discrete-time contour data defined only for $n = 1 \ldots N$ to functions which are defined for all time instances t:

$$x^*(t) = \sum_{k=1}^{K} c_k \xi_k(t)\,, \tag{1}$$

where the ξ_k are the new base functions and the coefficients c_k determine their weight. By choosing the number of bases K one can obtain a trade-off between smoothing of undesired fluctuations and retaining fine details. The coefficients c_k are determined in a minimum error sense. The calculation is controlled by a penalty parameter λ balancing fitting error and roughness of the curve. Similar to standard PCA, in the next step we calculate a new orthonormal basis $\varphi_u(t)$ using PCA in the functions domain. Assuming zero mean for $x^*(t)$ the projections of the smoothed input segment $x^*(t)$ onto this new basis, also termed Principal Components, is:

$$y_u = \int \varphi_u(t)\, x^*(t) dt \quad . \tag{2}$$

The function $x^*(t)$ is then reconstructed using the U basis functions:

$$\hat{x}^*(t) = \sum_{u=1}^{U} y_u \varphi_u(t)\,. \tag{3}$$

We retain the projections y_u as coefficients describing the contour segment. Due to the different feature rates in the acoustic (100 Hz) and visual (25 Hz) channel we used 30 and 10 points respectively for the linear interpolation of the segments on an

ISBN 978-3-8007-3640-9

© VDE VERLAG GMBH · Berlin · Offenbach

		silence	inside	all
Flat Start	non-adapted	0.71	0.47	0.53
	adapted	0.71	0.43	0.49
Segmented	non-adapted	0.47	0.23	0.28
	adapted	0.50	0.18	0.25
Onset Features	non-adapted	0.39	0.23	0.27
	adapted	0.23	0.19	0.20

Table 1: Alignment costs for the different training strategies with and without speaker adaptation of the HMM models evaluated on different segments.

equally spaced grid. The FPCA smoothing was done on the same grid using B-splines. All steps for the FPCA were performed with the Matlab toolbox retrieved from [32]. We retained 10 coefficients for each dimension for the acoustic features and 7 for the visual ones along the time axis. The FPCA transformations were learned using all the data of one speaker.

5 Results

We evaluate the different alignments in two different respects. First by comparing the alignments obtained from the HMMs to manual alignments, and secondly by measuring the performance of our word prominence detection framework with the different alignments.

5.1 Comparison to manual alignments

To evaluate the performance of the alignment we manually annotated 240 utterances from 8 speakers of our CoCo corpus. We measured the difference between these manual annotations and those obtained via the approaches detailed in Sec. 3. As cost function we used a piece-wise linear function where deviations below 20 ms were penalized with zero, above 50 ms with one and between 20 ms and 50 ms with linear costs increasing from 0 to 1. We measured the costs for three types of segments: only segments inside the utterance (inside), from the initial silence to the first word or from the last word to the final silence (silence) or all segments combined (all). The averaged costs with and without adaptation of the HMM models to the speaker are depicted in Table 1. As can be seen the initialization using the alignments provided from the Grid corpus yields much better results on our CoCo corpus than performing Flat Start. When combining the Segmented initialization with the Onset Features described in Sec. 3.1 the performance notably increases further. Yet as can be seen when looking on the results for the different segment classes this further increase in performance is mainly due to a better segmentation of the beginning and end of the utterance. The costs for the segmentation inside the utterance do not change or deteriorate slightly. Overall the additional adaptation to the speaker improves the segmentation for most of the cases.

5.2 Impact on word prominence detection

Ultimately we want to use the word segmentation for the detection of prominent words. Hence the most pertinent approach to assess the different alignments is a functional evaluation. For doing so we train a system to discriminate prominent from non-prominent words with the different alignments. We used a Support Vector Machine (SVM) with a Radial Basis Function Kernel using LibSVM [33] for this task. Via a grid search on the whole CoCo dataset we determined the penalty parameter of the error term C and the variance scaling factor of the basis function γ. Prior to the grid search the data was normalized to the range $[-1 \dots 1]$. With the found optimal parameters an SVM was trained on 75% of the data and tested on the remaining 25%. A 30 fold cross validation was run, in which the data set was always split such that an identical number of elements is taken from both classes. To establish the 30 sets a sampling with replacement strategy was applied. This process was performed individually for each speaker. As the alignments following the speaker adaptation showed higher precision we only used those for this functional evaluation.

	audio	video	A+V
Flat Start	81.6	70.9	79.9
Segmented	80.7	70.1	79.6
Onset Features	80.1	70.8	79.7

Table 2: Unweighted accuracy averaged over all 12 speakers for each of the 3 different alignments after adaptation to the speaker.

We investigate three different feature combinations, only acoustic features (compare Sec. 4.1, only visual features (compare Sec. 4.2) and a combination of acoustic and visual features. For all these features we combined the functionals and the FPCA contour models. Table 2 shows the unweighted accuracy averaged over all 12 speakers for each of the 3 different alignments after adaptation to the speaker.

As can be seen from Table 2 despite the very different accuracy of the alignments they all show very similar performance for the prominent vs. non-prominent word discrimination. Surprisingly the worst performing alignment, i. e. Flat Start, shows the best results. The video channel by itself already yields good accuracies for all three alignments. Yet when it is combined with the acoustic channel, the performance decreases slightly.

6 Discussion

We first showed that the different training strategies yield very different alignment accuracies. Using the alignments from the training corpus (Grid) during the training almost halved the alignment cost on the test corpus (CoCo). Adding features more targeted towards temporal segmentation to the standard MFCC further reduced the alignment cost prior to speaker adaptation by 20%. This shows that one can obtain better alignments with features tailored to the segmentation task than with standard MFCC. The cost after speaker adaptation did not change much in this case. From the results we also see that the inclusion of the alignments in the training has a much bigger effect than the speaker adaptation.

The functional evaluation of the segmentations revealed that even a much better segmentation does not improve the prominent word vs. non-prominent word discrimination. It actually slightly deteriorated the results. The reasons for this are not clear at the moment. One possibility is that the extraction of the functionals and contour models is very robust with respect to the alignment. Another possible explanation is the way acoustic prominence is realized. When making a word more prominent, i. e. hyperactiuclating it, speakers commonly make the neighboring words less prominent, i. e. they are hypoarticulated [34]. Also using features of the neighboring word increases the detection of the target word [6, 35]. Therefore, segmentations spanning beyond the actual word might improve the features. On the other hand such erroneous segmentations will also lead to the inverse case where the segmentation is shorter than the word. These two effects combined should hence not lead to an improvement in the discrimination accuracies.

We also investigated the impact of the visual features. There we saw that the visual features by themselves yield good discrimination accuracies and that they are also very robust with respect to the different alignments. On the other hand combining the acoustic and visual features did not improve the overall results. We observed this behavior previously and could in particular show that the reliability of the visual features varies a lot from speaker to speaker [18]. As a consequence there are some speakers where the visual information deteriorates results and others where it significantly improves the results.

7 Acknowledgments

We want to thank Petra Wagner, Britta Wrede and Heiko Wersing for fruitful discussions. Furthermore, we are very grateful

to Rujiao Yan for helping in setting up the visual processing as well to Venkatesh Kulkarni for developing the Voicing Detection. Many thanks to Mark Dunn for support with the cameras and the recording system as well to Mathias Franzius for support with tuning the SVMs and Andrea Schnall and Merikan Koyun for help in the data preparation. Special thanks go to our subjects for their patience and effort.

References

[1] D. Litman, J. Hirschberg, and M. Swerts, "Characterizing and predicting corrections in spoken dialogue systems," *Computational linguistics*, vol. 32, no. 3, pp. 417–438, 2006.

[2] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. EUROSPEECH*, ISCA, 2005.

[3] G. Levow, "Identifying local corrections in human-computer dialogue," in *Eighth Int. Conf. on Spoken Lang. Proc. (ICSLP)*, 2004.

[4] F. Tamburini, "Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system," in *INTERSPEECH*, 2003.

[5] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, no. 2, pp. 690–701, 2007.

[6] A. Rosenberg, *Automatic detection and classification of prosodic events*. PhD thesis, Columbia University, 2009.

[7] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am*, vol. 58, no. 4, pp. 880–883, 1975.

[8] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[9] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," 2013.

[10] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[11] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Applied Signal Process.*, vol. 11, pp. 1260–1273, 2002.

[12] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, "Audiovisual speech recognition with missing or unreliable data," in *Proc. Int. Conf. Auditory Visual Speech Process. (AVSP)*, 2009.

[13] T. Yoshida, K. Nakadai, and H. Okuno, "Automatic speech recognition improved by two-layered audio-visual integration for robot audition," in *Proc. 9th IEEE-RAS Int. Conf. on Humanoid Robots*, pp. 604–609, IEEE, 2009.

[14] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility," *Psychological Science*, vol. 15, no. 2, p. 133, 2004.

[15] J. Beskow, B. Granström, and D. House, "Visual correlates to prominence in several expressive modes," in *Proc. INTERSPEECH*, pp. 1272–1275, ISCA, 2006.

[16] M. Swerts and E. Krahmer, "Facial expression and prosodic prominence: Effects of modality and facial area," *Journal of Phonetics*, vol. 36, no. 2, pp. 219–238, 2008.

[17] M. Heckmann, "Audio-visual evaluation and detection of word prominence in a human-machine interaction scenario," in *Proc. INTERSPEECH*, (Portland, OR), ISCA, 2012.

[18] M. Heckmann, "Steps towards more natural human-machine interaction via audio-visual word prominence detection," in *submitted to 2nd Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, 2014.

[19] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, United Kingdom: Cambridge University, 1995.

[20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, pp. 2421–2424, 2006.

[21] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[22] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, vol. 53, no. 5, pp. 736 – 752, 2011.

[23] "Replaygain 1.0 specification." `http://wiki.hydrogenaudio.org/`.

[24] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proc. INTERSPEECH*, (Antwerp), pp. 2765–2768, 2007.

[25] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proc. INTERSPEECH*, vol. 2, p. 3, 2005.

[26] M. Heldner, "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish," *Journal of Phonetics*, vol. 31, no. 1, pp. 39–62, 2003.

[27] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[28] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

[29] J. Ramsay and B. Silverman, *Functional data analysis*. Springer, 2005.

[30] G. Michele, F. Torreira, and L. Boves, "Using FDA for investigating multidimensional dynamic phonetic contrasts," *Preprint submitted to Journal of Phonetics*, 2013.

[31] J. P. Arias, C. Busso, and N. B. Yoma, "Energy and f0 contour modeling with functional data analysis for emotional speech detection," in *Proc. INTERSPEECH*, (Lyon, FR), 2013.

[32] J. Ramsay, "Functions for functional data analysis in R, SPLUS and Matlab." `http://www.psych.mcgill.ca/misc/fda/`.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[34] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[35] A. Schnall and M. Heckmann, "Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario," in *Proc. INTERSPEECH*, (Singapore), 2014.