Steps Towards More Natural Human-Machine Interaction via Audio-Visual Word Prominence Detection

Martin Heckmann(✉)

Honda Research Institute Europe GmbH, 63073 Offenbach/Main, Germany
martin.heckmann@honda-ri.de

Abstract. We investigate how word prominence can be detected from the acoustic signal and movements of the speaker's head and mouth. Our research is based on a corpus with 12 English speakers which contains in addition to the speech signal also videos of the talker's head. To extract the word prominence information we use on one hand functionals calculated on the features and on the other hand Functional PCA (FPCA) to extract information from the contours. Combining the functionals and the contour information we obtain a discrimination accuracy between prominent and non-prominent words of 81 %. We show in particular that the visual channel is very informative for some speakers. Furthermore, we also introduce a system which extracts the prominence information online while a user is interacting with the system. The online system only uses acoustic information.

Keywords: Audio-visual · Prominence · Contour · FPCA · Online

1 Introduction

Spoken language is much more than the words we say. It comprises also information on the speaker's traits and states as personality and emotional state [9,36,37]. A truly natural interaction with an artificial agent, be it a virtual agent or a robot, requires that the agent is able to recognize and synthesize all of these aspects of human communication. Many efforts have been done to equip agents with the abilities to infer and produce different affective states [8,35]. In addition to its affective dimension the prosody of a sentence modulates its meaning (e. g. from a statement to a question) or the relevance the different words have for the speaker in the utterance (e. g. wide vs. narrow focus). The necessity of incorporating these dimensions in the speech synthesis process has well been accepted. Yet on the analysis side much less effort has been spent, in particular when looking on the multimodal aspects of prosody. Words which are strongly emphasized by the speaker and hence are perceived as very prominent by the listener frequently indicate a correction after a misunderstanding [41]. Endowing a machine with the capabilities to use this information is the target of our and previous research [26,27]. These words are also visually prominent, i. e. when

R. Böck et al. (Eds.): MA3HMI 2014 Workshop, LNAI 8757, pp. 15–24, 2015.
DOI: 10.1007/978-3-319-15557-9_2

only observing and not acoutically listening to the speaker [2,5,16,29,40]. Previously algorithms have been developed to detect these prominent words from the acoustic signal [26,27,33]. In [17] we presented an approach which also integrates information on the speaker's head and mouth movement to detect prominent words. In this paper we extend our approach on one hand by including a model for the representation of prosodic contours over time, namely Functional Principal Component Analysis (FPCA). On the other hand we also present an online system which is able to detect the prominent words in a real-time interaction with the user. For the moment the online system only relies on the acoustic channel.

In the next section we introduce the dataset we used for our experiments and training of the online system. We describe the different features extracted from the acoustic and visual channel and in particular the contour modeling in Sect. 3 Following this Sect. 4 will present the results of the classification experiments and Sect. 5 will discuss them. After that we will introduce the online system in Sect. 6. Then we will give a conclusion in Sect. 7.

2 Dataset

Our target scenario is the detection of corrected words via prosodic cues while a human is interacting with an agent. 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) and thereby making the corrected word highly prominent. The system interacted with the user via verbal feedback using the FESTIVAL speech synthesis system [6], pictures of a cartoon robot performing certain gestures, and visually highlighting different parts of the game. In total 16 native English speaking subjects were recorded [18]. 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 pixel and a frame rate of 25 Hz was used.

We trained HTK [43] on the Grid Corpus [12] followed by a speaker adaptation with a Maximum Likelihood Linear Regression (MLLR) step with a subsequent Maximum A-Posteriori (MAP) step to perform a forced alignment of the data. This forced alignment provides a temporal annotation of the data and is used in the following to determine the boundaries of the different words. Thereby we used a combination of RASTA-PLP and spectro-temporal HIST features [19] as this gave upon visual inspection better results than either of the feature sets alone or MFCC features.

Three human annotators annotated the recorded data with 4 levels of prominence for each word. We calculated the inter-annotator agreement with Fleiss' kappa κ. While doing so we binarized the annotations, i. e. only differentiating between prominent and non-prominent. We tested different binarizations and used the one where the agreement between all annotators was highest. Next, we calculated κ for each speaker individually. We then discarded all speakers where κ for the optimal binary annotation was below 0.5 ($0.4 < \kappa \leq 0.6$ is usually considered as moderate agreement). We have chosen such a rather low threshold to retain as many speakers as possible. This yields 12 speakers, 6 females and 6 males. 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 Features

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

3.1 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 [1], 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 [21]), 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 [24]. 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 $1.5 f_0$ [22].

3.2 Visual Features

To extract features from the visual channel we used the OpenCV library [7]. By its help we detected 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 from the

nose. 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.

3.3 Functionals

Functionals are commonly extracted from the (acoustic) features to detect prominent words [26,27,33] or perform prosodic analysis in general [15]. We extracted the mean, max, min, spread (max-min) and variance along the word. Hereby word boundaries were determined by the forced alignment. Prior to the calcualtion of the functionals and the contour modeling detailed in the next section we normalized the prosodic features by their utterance mean and calculated their first and second derivative (except for duration).

3.4 Contour Modeling via Functional Principal Component Analyis

To capture all the information in the feature and to be more tolerant against noise a more holistic representation than functionals based on contours is promising. Different approaches have been proposed to exploit this contour information as e. g. the extraction of plateaus [42], approximation with Legendre polynomials [23], stylizations of pitch contours with line segments [38], the calculation of the DCT from the contour [15] and in [18] we proposed to use a probabilistic parabola fitting to the contour. Very recently the Functional Principal Component Analysis (FPCA) has been proposed to discriminate emotional from neutral speech [4]. As the results obtained by the FPCA looked promising we opted to also apply it to the task of word prominence detection.

Functional data analysis is a branch of statistics which operates on curves or functions instead to data points [3,28,32]. The first step in FPCA is 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 $\varphi(t)$ 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:

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

Table 1. Unweighted accuracies in % averaged over all 12 speakers. See the text for the meaning of the feature name and modeling abbreviations. A+V represents audio and video features combined. The asterisk indicates values which are statistically significantly better than using the functionals only for the given feature combination ($\alpha = 5\,\%$).

	l	f_0	Nose	DCT	Audio	Video	A+V
Functionals	71.6	76.1	67.1	69.4	79.0	69.7	79.3
FPCA	71.6	76.6	67.0	68.4	79.7	71.2	78.9
Functionals+FPCA	72.8*	77.8*	67.7	70.2	80.7*	71.2*	80.4*

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. To calculate the FPCA we first linearly interpolated all segments on an equally spaced grid using B-splines. Due to the different feature rates in the acoustic (100 Hz) and visual (25 Hz) channel we used 30 and 10 points respectively. All steps for the FPCA were performed with the Matlab toolbox retrieved from [31]. We retained 10 coefficients for each dimension for the acoustic features and 7 for the visual ones along the time axis. The transformations were learned using all the data of one speaker.

4 Results

To discriminate prominent from non-prominent words we used a Support Vector Machine (SVM) with a Radial Basis Function Kernel implemented in LibSVM [11]. For each feature combination a grid search for C, the penalty parameter of the error term, and γ, the variance scaling factor of the basis function, was performed using the whole dataset. Prior to the grid search the data was normalized to the range $[-1 \ldots 1]$. With the found optimal parameters we trained an SVM on 75 % of the data and tested on the remaining 25 %. Hereby a 30 fold cross validation in which the data set was always split such that an identical number of elements is taken from both classes was run. To establish the 30 sets a sampling with replacement strategy was applied. This process was performed individually for each speaker.

As features we investigated loudness (l), f_0, the 2D nose position, the DCT calculated from the mouth region, the combination of all acoustic features, i. e. l, f_0, SE and duration, the combination of all visual features, i. e. nose and mouth DCT, and the combination of the acoustic and visual features. Table 1 shows the results averaged over all 12 speakers. For each of these feature sets we calculated

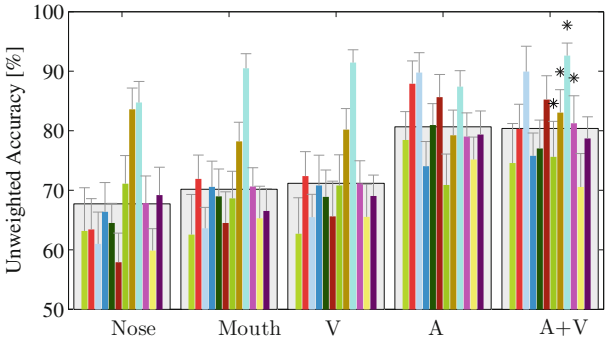

Fig. 1. Unweighted accuracies for different feature combinations (A represents results obtained from audio only, V from video only and A+V from the combination of audio and video). Each colored line indicates results for an individual speaker. The thin horizontal line on top visualizes the variance of the 30-fold cross-validation. An asterisk indicates a statistically significantly better result of the combination of A+V compared to A alone ($\alpha = 5\,\%$). The thick grey bar in the background depicts the mean over all speakers (Color figure online).

the unweighted accuracy when using only the functionals, the FPCA contour models or a combination of both.

As can be seen from Table 1 f_0 is clearly the strongest feature. The loudness l also performs quite well. For all feature sets the combination of the FPCA contour models with the funtionals yields equal or better results than either one alone. Cases where the difference is statistically significant are highlighted with an asterisk in Table 1.

In Fig. 1 results when using funtionals and FPCA are detailed for each speaker. As can be seen the nose and mouth DCT features perform well above chance level for all speakers. Figure 1 further shows that despite the decrease in performance on average when combining visual and acoustic features there is a strong gain from the visual features for some speakers.

5 Discussion

The results showed that the inclusion of prosodic contour models calculated via FPCA is in general beneficial for the detection of word prominence. They show in particular that not only the f_0 and energy contours carry important information but also the nose, resp. the head, movements and the evolution of the DCT coefficients calculated from the mouth region over time. As best results when looking on head movements (i. e. nose) only we obtain 67.7 % correct. When considering only the mouth movements (i. e. mouth DCT) we obtain 70.2 %. The best score we see when combining all visual features is 71.2 %. This is similar to the individual acoustic features but clearly inferior to the combination of all acoustic features. The overall best result we obtain with 80.7 % uses a combination of functionals and FPCA contour models of the acoustic features. A more

detailed analysis of the visual and audio-visual results revealed that there is a large variation from speaker to speaker. These variations seem to be very individual as we found in [18] that a grouping of the speakers based on the accuracies obtained with the different acoustic features (e. g. fundamental frequency, energy or duration) did not yield consistent groups. With the current data we see most notably that for some speakers the visual channel improves results yet for others it deteriorates them. The nose movement yields for all speakers accuracies well above chance level (57.9 % for the worst speaker and 67.7 % on average). For two speakers we see more than 80 % correct when looking on the nose movements only. It has previously been shown that the nose, resp. rigid head, movement differs between prominent and non-prominent words, even though there was no consensus if this is the case for all speakers [13,14,16]. In our previous work on the same dataset we also only saw accuarcies above chance level for some speakers [18]. Yet due to the improved visual feature extraction we now see it for all speakers and in particular are able to quantify the information content via our recognition experiments, at least to some extend. When we combine the nose movements with the mouth features we obtain 91.5 % correct for one individual speaker. This is also the one speaker where we observe very significant improvements from the combination of the acoustic and visual channel (87.4 % vs. 92.6 %). In total we see similar improvements from the combination of visual and acoustic features for 4 speakers. However, averaged over all speakers results are inferior to using the acoustic features alone. We think that with further improvements on the extraction of the visual features they will show beneficial in general when combined with acoustic features.

6 Online System

It is well known that users change their speaking style when talking to a machine as compared to when talking to a human [39]. Whereas they usually use a rich intonation when they talk to humans, they talk with a rather flat and monotonous voice to a machine. Currently we are integrating the different aspects we presented above in an online system. Such an online integration will allow us to investigate how users adapt to a system which is sensitive to prosodic variations. They might adapt to a speaking style as they would apply when talking to a human. Yet they might as well adapt to an exaggerated style which can better be decoded by a machine able to extract prosodic variations but not with the same aptness as a human.

So far our online system only extracts acoustic prosodic features. It comprises speech recognition, prosodic feature extraction, prosodic functionals calculation, word prominence labeling and visual feedback to the user. The speech recognition module decodes what the user has said and performs a word level temporal alignment which is required for the calculation of the prosodic functionals. We decided to use the HARK system developed at Honda Research Institute Japan together with Kyoto University to perform speech recognition [30]. The HARK system enables multi-channel speech enhancement and features the communication with the Julius speech recognition system via TCP/IP sockets [25].

We extended Julius with the possibility to communicate the recognition results and the temporal alignments also via TCP/IP to our RTBOS middelware which allows the component-based development of real-time systems [10]. We perform all remaining steps in our RTBOS framework which in particular allows a flexible distribution of processing modules to threads. For the extraction of the fundamental frequency we use the online implementation of our pitch tracking algorithm detailed in [20]. The extraction of the remaining acoustic features and the calculation of the functionals is straight forward. We have not yet integrated the contour modeling in the online system. For the word prominence labeling we also use in the online system the LibSVM [11].

7 Conclusion

In this paper we demonstrated on one hand how modeling the prosodic contours via FPCA can improve the detection of prominent words. On the other hand we also showed that for some speakers the visual channel can be efficiently used to detect prominent words. In particular we obtained from the head movements alone accuracies clearly above chance level for all speakers. Next we will extend the level of context available to the system by integrating the contour modeling with a decoding of the word sequence [34]. Furthermore, we will investigate how users adapt to a system capable of processing word prominence.

Acknowledgments. I want to thank Petra Wagner, Britta Wrede and Heiko Wersing for fruitful discussions. Furthermore, I am very grateful to Rujiao Yan and Samuel Kevin Ngouoko for helping in setting up the visual processing and the forced alignment, respectively 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, Paschalis Mikias and Merikan Koyun for help in the data preparation. Special thanks go to my subjects for their patience and effort.

References

1. Replaygain 1.0 specification. http://wiki.hydrogenaudio.org/
2. Al Moubayed, S., Beskow, J.: Effects of visual prominence cues on speech intelligibility. In: Proceedings of International Conference on Auditory Visual Speech Process. (AVSP), vol. 9, p. 16. ISCA (2009)
3. Arias, J.P., Busso, C., Yoma, N.B.: Energy and f0 contour modeling with functional data analysis for emotional speech detection. In: Proceedings of INTERSPEECH, Lyon, FR (2013)
4. Arias, J.P., Busso, C., Yoma, N.B.: Shape-based modeling of the fundamental frequency contour for emotion detection in speech. Comput. Speech Lang. **28**(1), 278–294 (2014)
5. Beskow, J., Granström, B., House, D.: Visual correlates to prominence in several expressive modes. In: Proceedings of INTERSPEECH, pp. 1272–1275. ISCA (2006)
6. Black, A., Taylor, P., Caley, R.: The festival speech synthesis system. Technical report (1998)

7. Bradski, G.: The openCV library. Dr. Dobb's J. Softw. Tools **25**, 122–125 (2000)
8. Buendia, A., Devillers, L.: From informative cooperative dialogues to long-term social relation with a robot. In: Mariani, J., Devillers, L., Garnier-Rizet, M., Rosset, S. (eds.) Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice, pp. 135–151. Springer, Heidelberg (2014)
9. Campbell, N.: On the use of nonverbal speech sounds in human communication. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 117–128. Springer, Heidelberg (2007)
10. Ceravola, A., Stein, M., Goerick, C.: Researching and developing a real-time infrastructure for intelligent systems - evolution of an integrated approach. Robot. Auton. Syst. **56**(1), 14–28 (2008)
11. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/ cjlin/libsvm
12. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**, 2421–2424 (2006)
13. Cvejic, E., Kim, J., Davis, C., Gibert, G.: Prosody for the eyes: quantifying visual prosody using guided principal component analysis. In: Proceedings of INTERSPEECH. ISCA (2010)
14. Dohen, M., Lœvenbruck, H., Harold, H., et al.: Visual correlates of prosodic contrastive focus in french: description and inter-speaker variability. In: Speech Prosody, Dresden, Germany (2006)
15. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of International Conference on Multimedia, pp. 1459–1462. ACM (2010)
16. Graf, H., Cosatto, E., Strom, V., Huang, F.: Visual prosody: facial movements accompanying speech. In: International Conference on Automatic Face and Gesture Recognition, pp. 396–401. IEEE (2002)
17. Heckmann, M.: Audio-visual evaluation and detection of word prominence in a human-machine interaction scenario. In: Proceedings of INTERSPEECH. ISCA, Portland (2012)
18. Heckmann, M.: Inter-speaker variability in audio-visual classification of word prominence. In: Proceedings of INTERSPEECH, Lyon, France (2013)
19. Heckmann, M., Domont, X., Joublin, F., Goerick, C.: A closer look on hierarchical spectro-temporal features (HIST). In: Proceedings of INTERSPEECH, Brisbane, Australia (2008)
20. Heckmann, M., Gläser, C., Vaz, M., Rodemann, T., Joublin, F., Goerick, C.: Listen to the parrot: demonstrating the quality of online pitch and formant extraction via feature-based resynthesis. In: Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Nice (2008)
21. Heckmann, M., Joublin, F., Goerick, C.: Combining rate and place information for robust pitch extraction. In: Proceedings of INTERSPEECH, pp. 2765–2768, Antwerp (2007)
22. Heldner, M.: On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish. J. Phonetics **31**(1), 39–62 (2003)
23. Jeon, J., Wang, W., Liu, Y.: N-best rescoring based on pitch-accent patterns. In: Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 732–741. Association for Computational Linguistics (2011)

24. Kristjansson, T., Deligne, S., Olsen, P.: Voicing features for robust speech detection. In: Proceedings of INTERSPEECH, vol. 2, p. 3 (2005)
25. Lee, A., Kawahara, T.: Recent development of open-source speech recognition engine julius. In: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 131–137 (2009)
26. Levow, G.: Identifying local corrections in human-computer dialogue. In: Eighth International Conference on Spoken Language Processing (ICSLP) (2004)
27. Litman, D., Hirschberg, J., Swerts, M.: Characterizing and predicting corrections in spoken dialogue systems. Comput. Linguist. **32**(3), 417–438 (2006)
28. Michele, G., Torreira, F., Boves, L.: Using FDA for investigating multidimensional dynamic phonetic contrasts. Preprint submitted to Journal of Phonetics (2013)
29. Munhall, K., Jones, J., Callan, D., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility. Psychol. Sci. **15**(2), 133 (2004)
30. Nakadai, K., Okuno, H., Nakajima, H., Hasegawa, Y., Tsujino, H.: An open source software system for robot audition hark and its evaluation. In: Proceedings of IEEE-RAS International Conference on Humanoid Robots (2008)
31. Ramsay, J.: Functions for functional data analysis in R, SPLUS and Matlab. http://www.psych.mcgill.ca/misc/fda/
32. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, New York (2005)
33. Rosenberg, A.: Automatic detection and classification of prosodic events. Ph.D. thesis, Columbia University (2009)
34. Schnall, A., Heckmann, M.: Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario. In: Proceedings of INTERSPEECH, Singapore (2014)
35. Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., Wöllmer, M.: Building autonomous sensitive artificial listeners. IEEE Trans. Affect. Comput. **3**(2), 165–183 (2012)
36. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: Paralinguistics in speech and languag-state-of-the-art and the challenge. Comput. Speech Lang. **27**(1), 4–39 (2013)
37. Shriberg, E.: Spontaneous speech: How people really talk and why engineers should care. In: Proceedings of EUROSPEECH. ISCA (2005)
38. Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A.: Modeling prosodic feature sequences for speaker recognition. Speech Commun. **46**(3), 455–472 (2005)
39. Shriberg, E., Stolcke, A., Hakkani-Tür, D.Z., Heck, L.P.: Learning when to listen: detecting system-addressed speech in human-human-computer dialog. In: Proceedings of INTERSPEECH (2012)
40. Swerts, M., Krahmer, E.: Facial expression and prosodic prominence: effects of modality and facial area. J. Phonetics **36**(2), 219–238 (2008)
41. Swerts, M., Litman, D., Hirschberg, J.: Corrections in spoken dialogue systems. In: Sixth International Conference on Spoken Language Processing (ICSLP). ISCA, Bejing (2000)
42. Wang, D., Narayanan, S.: An acoustic measure for word prominence in spontaneous speech. IEEE Trans. Audio Speech and Lang. Proc. **15**(2), 690–701 (2007)
43. Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University, Cambridge (1995)

http://www.springer.com/978-3-319-15556-2

Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction
Second International Workshop, MA3HMI 2014, Held in Conjunction with INTERSPEECH 2014, Singapore, Singapore, September 14, 2014, Revised Selected Papers
Böck, R.; Bonin, F.; Campbell, N.; Poppe, R. (Eds.)
2015, XII, 109 p. 29 illus., Softcover
ISBN: 978-3-319-15556-2