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

Inter-speaker variability in audio-visual classification of word prominence

Martin Ernst Heckmann

2013

Preprint:

This is an accepted article published in Proc. INTERSPEECH. The final authenticated version is available online at: https://doi.org/[DOI not available]

Inter-speaker variability in audio-visual classification of word prominence

Martin Heckmann

Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany

martin.heckmann@honda-ri.de

Abstract

In this paper we present results for the audio-visual discrimination of prominent from non-prominent words on a dataset with 16 speakers and more than 5000 utterances. We collected data in an experiment where users were interacting via speech in a small game, designed as a Wizard-of-Oz experiment, with a computer. Following misunderstandings of one single word of the system, users were instructed to correct this word using prosodic cues only. Hence we obtain a dataset which contains the same word with normal and with high prominence. We extract an extensive range of features from the acoustic and visual channel. Thereby we also introduce fundamental frequency curvature as a measure. The analysis shows that there is a large variation from speaker to speaker in respect to the discrimination accuracy between prominent and non-prominent words as well to which features yield the best results. In particular we show that the visual channel is very informative for many of the speakers and that overall the feature capturing the mouth shape is the best individual feature. Furthermore, we show that a combination of the acoustic and visual features improves the performance for many of the speakers.

Index Terms: prosody, prominence, visual, audio-visual, pitch curvature, polynomial approximation

1. Introduction

Current spoken dialog systems don't listen to the prosodic variations of speech even though it is well known that prosodic cues play a very important role in human communication [1]. Nevertheless, quite a few researchers have investigated how they could use prosodic cues to improve the human-machine dialog [2, 3, 4]. In general the inclusion of prosodic cues is quite difficult as they show not only a large variability from speaker to speaker but are also difficult to extract from the speech signal. The inclusion of visual information might be a route to alleviate these problems. Information on the movements of the speaker's mouth and face notably improves the accuracies of automatic speech recognition, particular in difficult situations [5, 6, 7, 8]. Humans are also able to use such visual information to extract prosodic cues [9, 10, 11, 12, 13]. Studies quantifying these visual prosodic cues have shown that they are mainly manifested in larger jaw opening, lip spreading and protrusion and to some extend to head movements [14, 15].

In [16] it was shown that speakers use prosodic cues to highlight corrections in a dialog with a machine and that these can be detected using prosodic cues. We extended this idea in [17] to the audio-visual discrimination of prominent from non-prominent words. In particular we showed that the performance can be improved by visual features extracted from the speaker's face without the use of additional visual markers. As visual features we used image transformations calculated on the mouth region of the speaker. For this paper we introduce two new acoustic features and significantly extended the dataset. The latter allows us to perform a more detailed analysis of the quality of the different features and their variation from speaker to speaker.

In the next section an overview on the recording of the data will be given. After that Section 3 describes the different features extracted from the acoustic and visual channel. Following this Section 4 will present the results of the classification experiments. Then we will discuss the results in Section 5 and give a conclusion in Section 6.

2. Dataset

For the recording of the data subjects interacted via speech in a Wizard of Oz experiment with a computer in a small game where they moved tiles to uncover a cartoon. With this playful setting we expected to obtain more natural speech, in particular regarding the prosody. 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. Verbal feedback was based on the FESTIVAL speech synthesis system [18]. 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. In total 16 subjects, 8 females and 8 males, eight speaking British English as their sole native language, three being bilingual British English/German, four speaking American English as their sole native language and one being bilingual American English/German were recorded. 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 [19] on the Grid Corpus [20] 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. Thereby we used a combination of RASTA-PLP and spectro-temporal HIST features [21] as this gave upon visual inspection better results then either of the feature sets alone or MFCC features.

For further processing those turns where the original utterance and a correction were available were selected. Overall we have 2683 turn pairs (original utterance + correction), i. e. on average $\approx$ 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. An analysis of acoustic

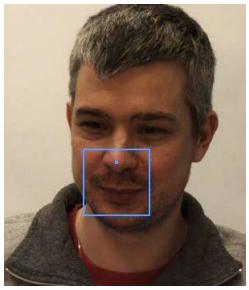

Figure 1: Image from recording after cropping to face region, nose detection, downsampling and highlighting of the mouth region.

features related to word prominence on a subset of the data in [17] showed that the words in the narrow focus condition were notably more prominent than in the broad focus condition .

3. Features

To extract word prominence from the acoustic signal we used on one hand features which are well known from the literature and on the other hand introduce two novel features: the maximal curvature of the fundamental frequency curve and the length of the gap before and after the word in question. For the visual channel we used the features we developed in [17]. More precisely we extracted the nose position in the image via the openCV library [22]. Starting from the nose position we used a fixed and for all speakers identical offset to determine the mouth region. After downsampling by a factor of 2 this yields an image of 100×100 pixels of the mouth region (compare Figure 1). On these images we calculated a two-dimensional Discrete Cosine Transform (DCT). Out of the 10000 coefficients per image we selected the 50 with the highest energy. We did this by calculating for each speaker separately the mean energy of all 10000 coefficients on a randomly selected subset of 10% of the data. Consequently we obtain 50 coefficients per frame to capture the mouth shape. We also investigated a two-dimensional Fast Fourier Transform (FFT) but as its performance was always inferior to that of the DCT we do not include the results here. From these features (except for duration) the mean value for each word was calculated and used in the subsequent analysis. The beginning and end of the word was taken from the forced alignment.

In the next sections we will detail the novel acoustic features. An overview over all the features we used is given in Table 1.

3.1. Gap Duration

The duration of the word is known to be a feature to signal word prominence. We refer to this as D_W. When looking at the data we observed that some speakers had the tendency to lengthen the gap between the prominent word and its predecessor and successor. We capture this by adding the gap before and after the word under investigation to obtain the feature D_G. The duration of these gaps is taken from the forced alignment of the data. Subsequently we combine D_W and D_G to D.

3.2. Fundamental Frequency Curvature

The perception of prominence of a word is strongly influenced by the presence and height of a fundamental frequency peak [23]. Yet in the determination of this peak a baseline has to be defined. As the fundamental frequency varies across a sentence this baseline is not obvious. Common approaches calculate some statistics of the fundamental frequency contour to capture this peak (e. g. max, min, mean, slope [24]) In order to capture this peak more robustly and to some extend more dependent on the local fundamental frequency context we introduce the maximal fundamental frequency curvature as a measure. The maximal curvature defines how pronounced the peak is. To calculate the curvature we first approximate the fundamental frequency (in the voiced part) of the word via a parabola $y = a_2x^2 + a_1x + a_0$. This parabola is fitted in a least squares sense. Outliers in this fit are removed with a RANSAC [25] like approach where we randomly fit different parabolas and select the one with the best trade-off between mean error and number of outliers:

$$\underset{n}{\operatorname{argmin}} \left[\bar{e}_{\text{fit}}(n)(1 + l(n))\right]$$

with $\bar{e}_{\text{fit}}(n)$ the mean error of the n-th approximation and $l(n)$ the corresponding number of outliers. Once we fitted the parabola we can determine the value and curvature

$$c = \frac{2a_2}{[1 + (2a_2x + a_1)^2]^{\frac{3}{2}}}$$

at the vertex $v = -a_1/(2a_2)$. For parabolas where the value at the vertex is a maximum this value defines our maximal fundamental frequency and maximal curvature value. For upwards oriented parabolas (i. e. a fundamental frequency valley) or if the vertex is outside the word boundaries we set the curvature to zero and the maximal fundamental frequency to the mean of the segment. In addition to these two features, i. e. the value at the vertex f_0^M and the curvature at the vertex f_0^c, we also extract $\bar{f}_0$, the mean fundamental frequency in the voiced region of the word relative to the mean of the utterance. In the following we use a combination of these three features as our fundamental frequency feature f_0. For the extraction of the fundamental frequency we use our previously developed algorithms for fundamental frequency extraction [26] and tracking [27].

e	energy of the word relative to the mean of the utterance
D_W	duration of the word
D_G	sum of the duration of the gap before and after the word (compare Sec. 3.1)
D	combination of D_W and D_G
f_0^M	maximal value of the fundamental frequency, i. e. value at vertex (compare Sec. 3.2)
f_0^c	maximal curvature of the fundamental frequency, i. e. curvature at vertex
$\bar{f}_0$	mean fundamental frequency of the word relative to the mean of the utterance
f_0	combination of f_0^M, f_0^C and $\bar{f}_0$
DCT	50 DCT coefficients extracted from the mouth region with the highest energy
y	nose y position relative to the mean of the utterance
$y\Delta$	nose y velocity

Table 1: Features used to extract word prominence. For details see the text, in particular Sec. 3.1 and 3.2.

4. Results

To discriminate prominent from non-prominent words a Support Vector Machine (SVM) with a Radial Basis Function Ker-

D_W	D_G	D	e	SE	f_0^M	f_0^c	$\bar{f}_0$	f_0	DCT	y	$y\Delta$	e-D	e-D-SE	e-D-SE DCT	e-D-SE-f_0	e-D-SE-f_0 DCT
62.8	59.0	64.7	58.6	58.8	65.2	64.8	62.6	68.2	66.9	54.9	57.5	69.5	71.6	73.7	74.9	74.9

Table 2: Classification scores in % averaged over all 16 speakers. See table 1 for the meaning of the feature name abbreviations.

nel was trained using LibSVM [28]. 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 $[-1 \dots 1]$. With the found optimal parameters an SVM was trained 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.

In Table 2 the results of the individual features averaged over all 16 speakers are displayed. When comparing the score for D_W, i. e. the duration of the word, D_G, i. e. the duration of the gap before and after the word, and the combination of both D one can see that combining the two values into one feature notably improves the performance. Similarly, the combination feature f_0 consisting of the fundamental frequency value at the vertex of the parabola approximation f_0^{Max} with the curvature at the vertex f_0^{Curv} and the mean of the fundamental frequency in the word relative to the utterance $\bar{f}_0$ also leads to a significant improvement. For these reasons in the following we will only use the combined features D and f_0. One also notices that the DCT feature is quite strong. In fact only the (combined) f_0 feature is stronger. This is a clear indication that the mouth shape conveys a significant amount of information on the prominence of the word. The nose position and speed on the other hand yield only little information.

In Fig. 2 the results for each individual speaker are given. As can be seen the quality of the different features for the different speakers varies a lot. Fundamental frequency is the strongest feature for 6 speakers (A, C, D, E, G, I). Duration only for 4 speakers (B, L, M, P) and relative energy for none. Spectral emphasis is a strong cue for speakers F and G, yet not the strongest. On the other hand the DCT feature is also for 6 speakers the strongest feature (F, H, J, K, N, O). The nose y position and in particular the velocity show high scores for speakers F, I and P.

Next we want to investigate how the combination of these different features behaves. When looking at Table 2 and Fig. 3 we see that the combination of energy and duration yields to a notable improvement. Adding also spectral emphasis improves further, up to 71.6% correct. If we now also add the DCT features we further improve to 73.7% correct. The combination of energy, duration, spectral emphasis and fundamental frequency is the best audio-only combination with 74.9% correct. Adding to this combination also the DCT feature does not increase the average results any further and also yields 74.9% correct. A look at Fig. 3 reveals again that the performance for the different speakers is quite varied. In the combination of energy, duration and spectral emphasis adding the DCT improves results for many speakers (those where the improvement is statistically significant, evaluated with a t-test with $\alpha = 5\%$, are marked with an asterisk). These are speakers C, D, F, H, J, L, N, O and P, i. e. 9 speakers. Hence almost all speakers where the DCT was the strongest individual feature and some more. When also the fundamental frequency is added there is only an improvement for 5 speakers (C, D, H, I, and J).

Finally we want to take a look on the correlation of the features with each other and the class labels in Fig 4. In particular

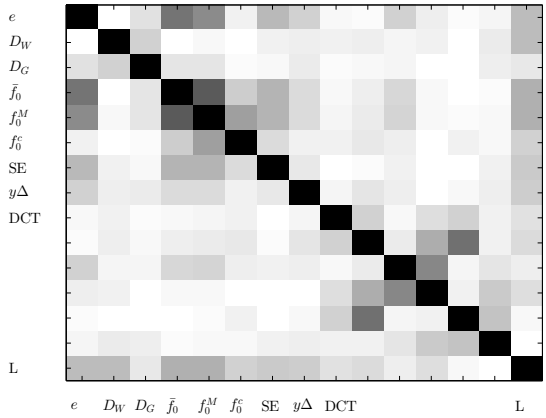

Figure 4: Covariance matrix of the different features with each other and the labels (L) for speakers F, H and J. For the DCT the coefficients are ordered with underlying y and x frequencies as (0,0), (0,1), (1,0), (1,1), (0,2), (2,0).

we are interested in the correlation of the visual features with the acoustic features. For this reason we only used the data from speakers F, H and J as they had particularly high scores for the visual features. In case of the DCT we used the five coefficients with the lowest frequencies. As can readily be seen we replicate previous findings that energy and fundamental frequency are highly correlated [29]. Furthermore, we can see that f_0^c, the curvature of the fundamental frequency, is not strongly correlated to f_0^M and $\bar{f}_0$, the maximal and mean value, respectively. Hence the curvature yields novel information. Similar D_G, the duration of the gap, is only rather weakly correlated with D_W, the duration of the word and therefore also yields novel information. Regarding the visual features the situation is more difficult. As well the nose y velocity as the DCT lowest y frequency coefficient (0,1) are correlated with the energy. The visual features show only little correlation with duration. A look at the correlation of the DCT coefficients with the class labels shows that mainly the mean energy (0,0), and the lowest x (0,1), and y (1,0) frequencies as well as the lowest diagonal frequency (1,1) show correlations. The correlations for the higher frequencies (0,2) and (2,0) are very low.

5. Discussion

We first observed that the newly introduced features, i. e. the duration of the gap before and after the word as well as the approximation of the fundamental frequency with a parabola and the calculation of the value and curvature at the vertex, yielded to a notable improvement of the classification scores. This is substantiated by a feature correlation analysis which indicated that they contribute novel information.

The detailed analysis of the results showed that there is a large variation in classification accuracy from one speaker to the other. The best score for the best speaker (J) is 92% correct and for the worst (D) it is 66%. Furthermore, we observed that also the strength of the different features is quite different from speaker to speaker. For most speakers either duration, fundamental frequency or the visible articulators are the most

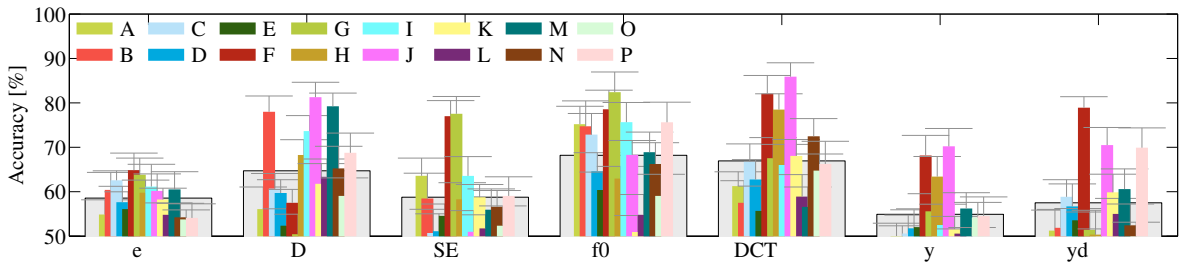

Figure 2: Discrimination accuracies for different features for each speaker (A ... P). The grey bars in the background visualizes the average over all speakers for a given feature. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. See Table 1 for an explanation of the abbreviations.

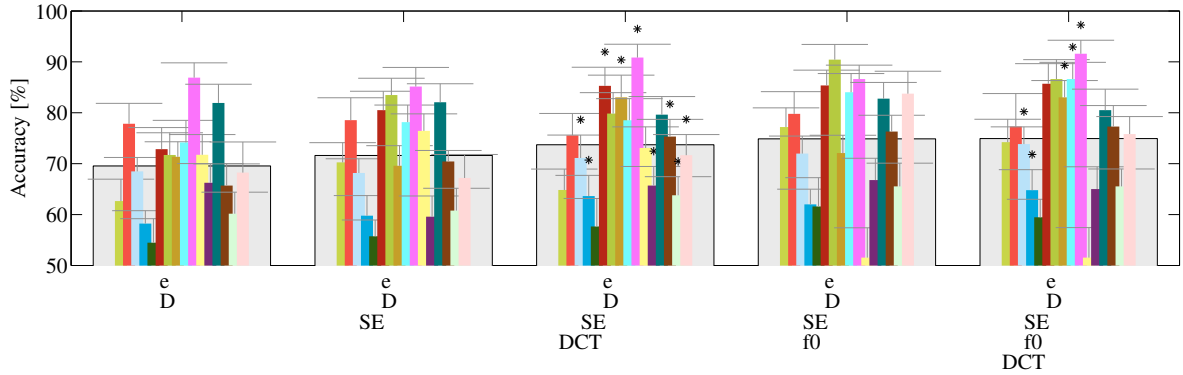

Figure 3: Discrimination accuracies for different feature combinations for each speaker. The grey bars in the background visualizes the average over all speakers for a given feature combination. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. The asterisk marks results where the combination of acoustic and visual features are statistically significantly better than the corresponding acoustic only results. See Table 1 for an explanation of the abbreviations.

informative. We see a trend that the American subjects (F, K, L, M, P) used duration to a lesser extend.

In our view the most surprising aspect is that the visual modality is for some speakers very informative with scores of $\approx 80\%$ correct. For the moment we do not really know what the main cue is which yields these good classification rates in the visual channel. The feature correlation analysis indicates that it is not a durational cue, e. g. longer opening of the mouth, they capture. This is supported by the fact that two speakers (F, H) who show very good scores for the DCT have only low scores for duration. We suspect that it is the visible hyper articulation for prominent words [14], i. e. a larger opening of the mouth and a wider spreading of the lips, which is responsible for the good classification scores. The correlation analysis supported this to some extend as mainly the DCT coefficients capturing the mouth opening and spreading show a correlation with the class labels.

A limitation in the current setup is that the extraction of the visual features is much less reliable than the extraction of the acoustic features. In addition to some imprecisions in the nose tracking quite a few speakers also tilted their head frequently which leads to a rotation of the mouth region. This then dramatically changes the DCT feature extraction. Hence we expect improvements for the visual feature extraction once this head rotation is compensated.

We could also see that some speakers use head movements (reflected by the nose position and velocity) to signal prominence. In particular using only the nose velocity yielded more than 70% correct rates for speakers F, J and P. This is in line with previous findings which in particular found that not all speakers behave like this [9, 14, 15]. Yet even for these speakers the recognition scores did not improve when the nose feature was either combined with the DCT features or the acoustic features.

6. Conclusion

We collected via a small Wizard-of-Oz game with a computer data where subjects were uttering words with normal and high prominence. During the game we made audio and video recordings. We then extracted acoustic and visual features and trained an SVM classifier to discriminate the normal from the highly prominent words. The results showed that there is a large difference between the different speakers. This comprises as well a large variation between the recognition scores we obtained as well as which features yield the best scores. In particular we could show that the visual channel contains a lot of information for the discrimination of prominent from non prominent words by a machine. For several speakers the combination of the acoustic and the visual channel significantly improved the recognition scores. Next steps include a correction of head tilt motions which dramatically compromise our visual feature extraction process and the move from a classification to a detection task.

7. 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. 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 Merikan Koyun for help in the data preparation. Special thanks go to my subjects for their patience and effort.

8. References

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

[2] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 5, pp. 519–532, 2000.

[3] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C.V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[4] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, no. 1-2, pp. 155–175, 2004.

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

[6] 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.

[7] 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.

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

[9] H.P. Graf, E. Cosatto, V. Strom, and F.J. Huang, "Visual prosody: Facial movements accompanying speech," in *Int. Conf. on Automatic Face and Gesture Recognition*. IEEE, 2002, pp. 396–401.

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

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

[12] 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.

[13] S. Al Moubayed and J. Beskow, "Effects of visual prominence cues on speech intelligibility," in *Proc. Int. Conf. Auditory Visual Speech Process. (AVSP)*. ISCA, 2009, vol. 9, p. 16.

[14] M. Dohen, H. Lœvenbruck, H. Harold, et al., "Visual correlates of prosodic contrastive focus in french: Description and inter-speaker variability," in *Speech Prosody*, Dresden, Germany, 2006.

[15] E. Cvejic, J. Kim, C. Davis, and G. Gibert, "Prosody for the eyes: Quantifying visual prosody using guided principal component analysis," in *Proc. INTERSPEECH*. ISCA, 2010.

[16] G.A. Levow, "Identifying local corrections in human-computer dialogue," in *Eighth International Conference on Spoken Language Processing*, 2004.

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

[18] A.W. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," Tech. Rep., 1998.

[19] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University, Cambridge, United Kingdom, 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, 2006.

[21] 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, Perceptual and Statistical Audition.

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

[23] Carlos Gussenhoven, Bruno H Repp, A Rietveld, Hans Hardy Rump, and Jacques Terken, "The perceptual prominence of fundamental frequency peaks," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 3009–3022, 1997.

[24] Andrew Rosenberg, "Autobi-a tool for automatic tobi annotation," in *Proc. INTERSPEECH*. ISCA, 2010.

[25] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

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

[27] M. Heckmann, C. Gläser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS)*, Nice, 2008.

[28] Chih-Chung Chang and Chih-Jen 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`.

[29] Andrew Rosenberg and Julia Hirschberg, "On the correlation between energy and pitch accent in read english speech," in *Ninth Int. Conf. Spoken Language Proc.*, 2006.