Confirmation detection in human-agent interaction using non-lexical speech cues

Mara Brandt, Britta Wrede, Franz Kummert and Lars Schillingmann
Cluster of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University, 33615 Bielefeld, Germany
{mbrandt, bwrede, franz, lschilli} at techfak.uni-bielefeld.de

Abstract

Even if only the acoustic channel is considered, human communication is highly multi-modal. Non-lexical cues provide a variety of information such as emotion or agreement. The ability to process such cues is highly relevant for spoken dialog systems, especially in assistance systems. In this paper, we focus on the recognition of non-lexical confirmations such as "mhm", as they enhance the system's ability to accurately interpret human intent in natural communication. We implemented and evaluated a system for online detection of non-lexical confirmations. The architecture uses a Support Vector Machine to detect confirmations based on acoustic features. In a systematic comparison, several feature sets were evaluated for their performance on a corpus of human-agent interaction in a setting with naive users including elderly and cognitively impaired people. Our results show that using stacked formants as features yield an accuracy of 84% outperforming regular formants and MFCC or pitch based features for online classification.

1 Introduction

In human-machine interaction it is important to provide an intuitive interface for the users that allows them to make free use of the modality space. Among humans, speech is one of the most important modalities to communicate information, although non-verbal modalities such as gaze, gesture or action have been shown to be highly relevant for establishing common ground as well.

Speech contains discourse particles or interjections which are important markers about the speaker's attitude (Andersen and Thorstein 2000). These particles, that can be part of an utterance or also standalone (interjections), help to ground not only propositional meaning but also convey epistemic states (Fetzer and Fischer 2007). Epistemic states pertain to the attitude of a speaker towards the information i.e. whether the speaker believes that the propositional content of an utterance (the own or the interlocutor's) is new and surprising or is already grounded, whether the speaker believes that the information is correct etc. This information is highly relevant also in HCI which still tends to be quite brittle with respect to grounding. It is important to make use of these rather subtle cues to infer the user's attitude towards the current interaction. It is therefore important for a human-agent interaction to change the dialog structure according to the perceived internal state of the user by slowing down or repeating if uncertainty is perceived, or by continuing if the user is confirming.

However, discourse particles/markers or interjections have certain characteristics that render them difficult for automatic recognition with standard ASR approaches: For one, their use and surface structure is highly variable between different speakers (Bell, Boye, and Gustafson 2001). Second, discourse particles are often characterized by stylized intonations, i.e. significantly different intonation patterns than "normal" speech (Gibbon and Sassen 1997). Indeed, it has been shown that extreme values for prosodic features yield a much higher word error rate in ASR systems (Goldwater, Jurafsky, and Manning 2010) making it difficult to recognize the lexical units of discourse particles. But also, the meaning of discourse particles depends on the underlying intonation (Gibbon and Sassen 1997). However, standard approaches to ASR do explicitly not take prosodic information into account.

In order to understand the meaning of the discourse particle, it is thus necessary to develop new approaches and investigate their acoustic nature.

One important feedback signal in dialogs is positive acknowledgment which indicates that the listener is still hearing and understanding what is being said. These feedback signals are often called "filled-pauses" and contain generally non-lexical acoustic units such as "mhm" or "aha". It has been shown that this feedback can be used by interaction partners to infer the listener's meta-cognitive state (Brennan and Williams 1995).

While the phonetic realizations may be variable, it has been shown that their prosodic cues remain very stable with a very slowly and smoothly declining F0 (Tsiaras, Panagiotakis, and Stylianou 2009). More specifically, fillers have a flat pitch, which lies in the median of pitch of the user across all his utterances (Garg and Ward 2006). Also, filled-pauses show a very specific articulation in that the articulators do not change their positions, yielding very stable formants and minimal coarticulation effects (Audhkhasi et al. 2009). This is acoustically reflected in small fundamental frequency transitions and small spectral envelope deformations (Goto, Itou, and Hayamizu 1999).

Figure 1: Interaction with the virtual agent "BILLIE"

2 Dataset

2.1 Scenario

Our research is part of the KOMPASS project (Yaghoubzadeh, Buschmeier, and Kopp 2015). In this project, a virtual agent "BILLIE" is developed to help elderly and cognitively impaired people to plan and structure their daily activities, get reminders and suggestions for possible activities. The users interact naturally with the system to enter their appointments, therefore the system needs to understand natural language inputs and react to feedback. In addition to visual cues for understanding and confirmation, e.g. nodding, it is important for the dialog system to detect non-lexical confirmations like "mhm", because the automatic speech recognition (ASR) typically doesn't recognize them.

2.2 User Study

As part of the KOMPASS project a user study with participants of the intended user groups, elderly and cognitively impaired people, was conducted. The study was performed as a Wizard of Oz experiment (Kelley 1984). 52 participants, consisting of 18 elderly (f: 14, m: 4), 18 cognitively impaired (f: 10, m: 8) and 16 students (f: 10, m: 6) with German as their first language, interacted with "BILLIE" and planned their daily activities for one week (Fig. 1). The participants only got instructions to enter their appointments naturally in German without being instructed to use any special commands or phrases, resulting in natural communication with the system.

2.3 Annotations

The KOMPASS WOZ1 corpus was annotated automatically and manually. Voice activity detection (VAD) was used to divide the speech signal into segments of continuous speech. All segments are automatically annotated as regular utterances, unless they contain non-lexical confirmations, which were manually annotated. The distribution of regular utterances and non-lexical confirmations as well as the subsets used for this evaluation can be seen in Tab. 1. The regular utterances contain 394 manually annotated filled-pauses,

set	#participants (f, m)	#all segments	#confirmations
WOZ1 data	52 (f: 34, m: 18)	5385	129
Training set	17 (f: 14, m: 3)	1885	87
Test set	4 (f: 3, m: 1)	415	42

Table 1: WOZ1 corpus segment distribution and used subsets for training with cross-validation and testing

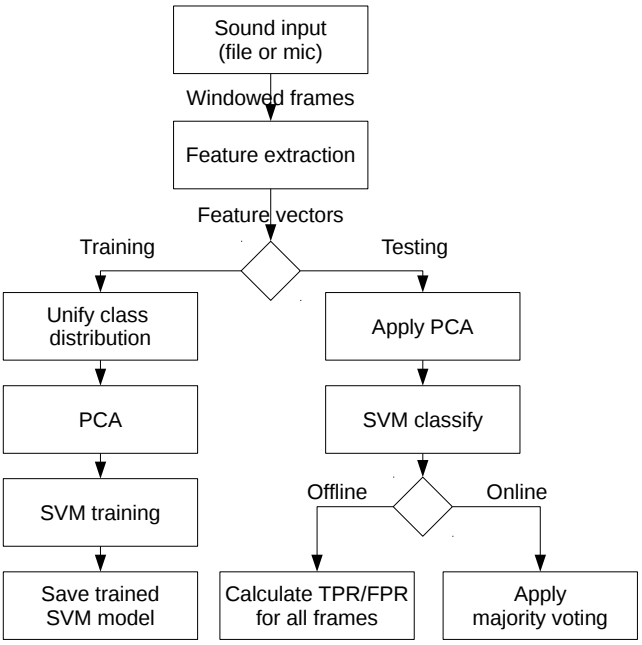


Figure 2: System architecture

e.g. elongations and fillers, some of them similar to confirmations, e.g. "hmm", which can lead to false-positives in the detection of non-lexical confirmations.

3 Non-Lexical Confirmation Detection System

3.1 Architecture

We developed a system for the recognition of non-lexical confirmations that can handle sound input from files or microphone recording. The architecture of this system is shown in Fig. 2. To support both offline and online processing with mostly the same algorithms, the software consists of modules that can be used in both modes. The input, using either sound files or a microphone as source, is chunked into overlapping frames of 25ms with a frame shift of 10ms. After that, the frames are windowed with a Blackman-Harris window (Harris 1978) for the MFCC related feature sets or a Hann window (Blackman and Tukey 1959) for formant and pitch based feature sets and the different selected features are extracted frame by frame. Principal Component Analysis (PCA) (Pearson FRS 1901) is performed to reduce the dimensionality of the feature vectors of feature sets with stacked features or derivatives except for stacked formants, which is necessary because of the high dimensionality of the stacked features and the small amount of data. In train-

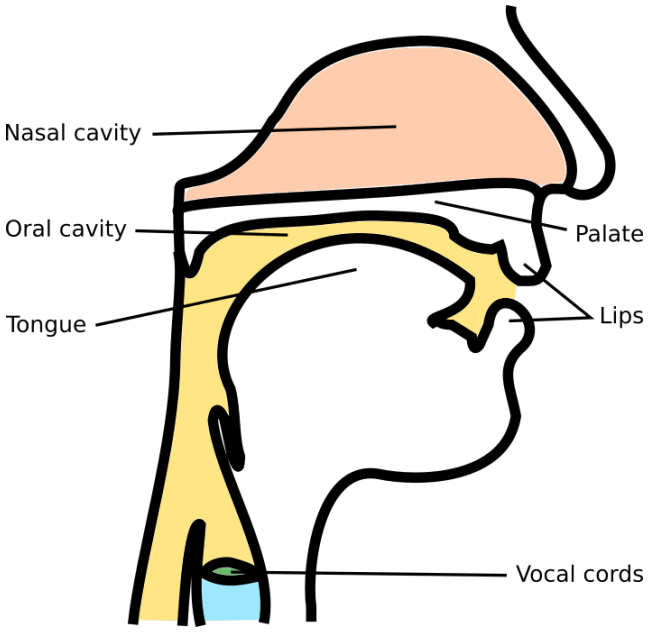


Figure 3: Source-filter model: vocal cords and vocal tract based on (Philippsen and Wrede 2017)

ing mode, a Support Vector Machine (SVM) (Vapnik 1995) is trained as described in Sec. 3.3 and the trained model is serialized for later classification tasks. For the classification mode, the same steps are required, but instead of the SVM training, the sound input is classified frame by frame with the deserialized trained SVM model. The classification results are calculated according to the description of offline and online classification in Sec. 3.4.

3.2 Feature Extraction/Selection

As argued above, different features may be suited for recognizing filled pauses. On the one hand, the pitch contour appears to be very salient, thus F0 becomes an interesting feature. On the other hand, the vocal tract (Fig. 3) remains stable, which would be reflected in the formants and the MFCCs. The different features are described in this section and an overview of the resulting feature vectors with corresponding sizes are shown in Tab. 3.

Mel-Frequency Cepstral Coefficients Mel-frequency cepstral coefficients (MFCCs) are features common in speech recognition. They reflect properties of the vocal tract during speech production and mimic human perception of speech. The coefficients are designed to mitigate speaker dependent characteristics. MFCCs are extracted with the Essentia framework (Bogdanov et al. 2013). For this the sound signal is windowed with a Blackman-Harris window and the spectrum of this window is computed. After that, the first 13 MFCCs in the frequency range from 20 Hz to 7800 Hz are calculated with 40 mel-bands in the filter.

Differentiation To capture the salient articulation we also calculated the first and second derivatives of the MFCCs (Δ and $\Delta\Delta$). For this the polynomial filter introduced by Savitzky and Golay (Savitzky and Golay 1964) is used which combines differentiation and smoothing. The general formula for this filter is shown in Equ. 1, where n is the filter length, a_i are the coefficients and h is the normalization factor.

$$y_t = \frac{1}{h} \sum_{i=-\frac{n-1}{2}}^{\frac{n-1}{2}} a_i x_{t+i} \quad (1)$$

Savitzky and Golay provide coefficients to use for the calculation of the derivatives (Tab. 2).

derivative	filter length	coefficients	h (normalization factor)
first	7	-3, -2, -1, 0, 1, 2, 3	28
second	7	5, 0, -3, -4, -3, 0, 5	42

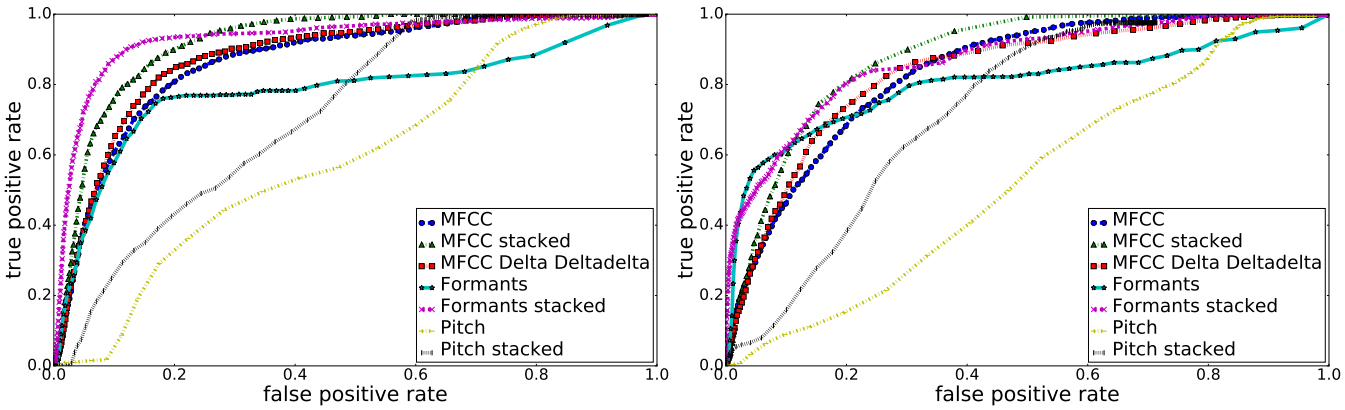
Table 2: Savitzky-Golay filter coefficients

Stacked MFCCs Stacked MFCCs are another way to model the context and dynamics of MFCCs and can outperform MFCC derivatives (Heck et al. 2013). Instead of calculating the derivatives, the 13 MFCCs of adjacent frames are stacked to form a single feature vector. 15 stacked frames, resulting in a 195-dimensional feature vector, yield the best results in terms of true positive and false positive rate for non-lexical confirmation recognition.

Formants As formants are directly correlated with movements of the vocal tract, they should be able to provide good features for filled-pause detection (see Sec. 1). Non-lexical confirmations are very similar to some filled-pauses, so formants can be used for non-lexical confirmation detection. The linear predictive coding (LPC) algorithm of the Essentia framework is used to calculate linear predictive coefficients of the order 12. These coefficients are used to calculate the polynomial roots using the Eigen3 PolynomialSolver algorithm (Guennebaud, Jacob, and others 2010). Subsequently, the roots are fixed into the unit circle. The first two formant frequencies which show the configuration of the vocal tract are calculated from these fixed roots. To measure the stability of the formants, the standard deviation of each formant over 15 frames is calculated and added as features.

Stacked Formants The idea of stacked MFCCs to model the dynamics of the signal over time can also be applied to formants. The 2 formants calculated per frame are therefore stacked over 15 frames to form one 30-dimensional feature vector.

Pitch Pitch is a feature that measures the frequency of the vocal cord vibrations. It is calculated using the PitchYinFFT algorithm (Brossier 2007) of the Essentia framework, which is an optimized version of the Yin algorithm calculated in the



(a) ROC plot of the first test set

(b) ROC plot of the second test set

Figure 4: ROC curves of classifiers with different feature sets

frequency domain. The input is therefore windowed with a Hann window.

Stacked Pitch Corresponding to the approach with MFCCs and Formants, the calculated pitch values over 15 frames are stacked to form one feature vector.

Principal Component Analysis Principal Component Analysis (PCA) is applied to the feature vectors of feature sets with stacked features or derivatives, except for stacked formants, to reduce dimensionality and to transform the features into linearly uncorrelated variables that describe the largest possible variances in the data. The algorithm used is the vector_normalizer_pca from the dlib library (King 2009). When the PCA is performed, the feature vectors are normalized automatically and no additional normalization is necessary. The PCA parameter ϵ controls the size of the transformed feature vector. A value $0 < \epsilon <= 1$ can be chosen and large values result in larger feature vectors. For our evaluation, we chose a value of 0.95 to maintain most of the information contained in the features.

Feature combination	Feature vector size
MFCCs	13
MFCCs Delta Deltadelta	39
Stacked MFCCs (15 frames)	195
SD of formants (15 frames)	2
Stacked formants (15 frames)	30
Pitch	1
Stacked pitch (15 frames)	15

Table 3: Tested feature combinations with corresponding feature vector sizes (without PCA)

3.3 Model estimation

A Support Vector Machine (SVM) with radial basis function (RBF) kernel is used for the classification of non-lexical confirmations vs. other speech. Therefore, the svm_c_trainer algorithm of the dlib library is used in the implementation. For the SVM training a feature vector of the selected features for each frame of a segment in the training set is calculated. All feature vectors are collected and the features are normalized as part of the PCA transformation. The normalized feature vectors are then used to train the SVM model.

3.4 Classification

Classification can be performed in offline and online mode. In offline mode, the classification results for each frame are stored and finally combined to calculate accuracy and true/false positive rates. For online classification, each frame is classified and the classification results are added as 1 for confirmations and -1 for other utterances to a rolling mean over 5 frames, which is used to implement a simple majority vote. If a specified majority threshold is reached the current VAD segment is classified as an utterance containing a non-lexical confirmation.

4 Evaluation

4.1 KOMPASS WOZ1 Data Preparation

The KOMPASS WOZ1 data is prepared for SVM training and offline classification by extracting all VAD segments from the sound files. If the interval contains a manual annotation of a non-lexical confirmation, the segment is added as belonging to the non-lexical confirmation class. A script is used to split all segments in training and test sets. Users without non-lexical confirmation utterances are excluded. We made sure to assign all utterances from one speaker to one of the sets only in order to achieve a speaker independent test set-up. 70 percent of the KOMPASS WOZ1 data are used for training, while the remaining 30 percent are used as test data (see Tab. 1). A leave-one-user-out cross-validation is performed on the training set. For each fold, all utterances of one user are used as the test set, while the remaining users are used for training. Because segments of non-lexical confirmations are usually shorter than other utterances, an additional step to maintain a uniform distribution of frames

feature set	feature vector size	cross-validation result	TPR (%)	FPR (%)	ROC AUC
MFCCs	13	76.5 - 81.4	82.7	20.2	0.87
MFCCs Delta Deltadelta	22	80.0 - 85.7	85.7	21.4	0.88
Stacked MFCCs	58	**91.0 - 96.8**	82.8	10.0	**0.92**
Formants	2	73.8 - 78.5	77.1	27.8	0.78
Stacked Formants	30	83.3 - 88.1	91.9	16.8	**0.93**
Pitch	1	11.7 - 17.6	99.8	92.0	0.60
Stacked Pitch	11	37.9 - 43.8	99.1	65.1	0.71

Table 4: Evaluation results: Cross-validation shows the stable performance of stacked MFCCs, but stacked formants achieve the highest ROC AUC for the test set

belonging to each class has to be performed. Feature vectors of other utterances are discarded prior to SVM training to compensate for the small number of frames belonging to non-lexical confirmations. The test set contains a realistic subset of unevenly distributed utterances of both classes and all frames of these utterances are classified without balancing the uneven distribution of feature vectors of both classes.

4.2 Parameter Optimization

Feature combination	C	ϵ	γ
MFCCs	1	0.5	0.005
MFCCs Delta Deltadelta	1	0.1	0.005
Stacked MFCCs (15 frames)	1	0.5	0.005
SD of formants (15 frames)	5	0.005	0.05
Stacked formants (15 frames)	1	0.5	0.05
Pitch	5	0.005	0.05
Stacked pitch (15 frames)	5	0.5	0.05

Table 5: Best SVM parameters found with grid search for each feature set

Grid search was used to optimize the SVM parameters C, ϵ and γ for the RBF kernel. The parameters were tested in the ranges $C \in \{1, 5\}$, $\epsilon \in \{0.005, 0.05, 0.1, 0.5\}$ and $\gamma \in \{0.005, 0.05\}$. The best results for each feature set are shown in Tab. 5.

4.3 Results on the KOMPASS WOZ1 Data

The system for non-lexical confirmation detection was tested on the KOMPASS WOZ1 data. Seven different feature sets were evaluated: MFCCs, MFCCs + first and second derivative (Δ, $\Delta\Delta$), stacked MFCCs, formants, stacked formants, pitch and stacked pitch. Grid search was performed as described in Sec. 4.2 for parameter optimization. Before the SVM was trained, a leave-one-user-out cross-validation was performed (see Sec. 4.1). To evaluate the performance of the trained models, the sum of the accuracy value weighted with the number of non-lexical confirmations for each fold was calculated.

Fig. 4 shows the Receiver Operating Characteristic (ROC) curves of the seven classifiers with different feature sets, that were evaluated on different test sets. For the first test set, the stacked formants can outperform all other feature sets with an area under the curve (AUC) of 0.93. In comparison, the standard deviation of the formants achieve an AUC of 0.78, which is even below all of the MFCC related feature sets. The feature vectors consisting of 13 MFCCs result in classification results with an AUC of 0.87 and adding first and second derivative only slightly improves the result (AUC of 0.88). Stacking of the MFCCs raises the AUC to 0.92 , but stays below the value of stacked formants. Using pitch as a single feature results in a nearly diagonal ROC curve (AUC of 0.60), which corresponds to classification results near chance level. Stacking the pitch values to feature vectors over 15 frames only slightly improves the results (AUC of 0.71). The results with the second test set show, that the performance of the formant related feature sets is not stable, while the results of MFCC related and pitch related feature sets remain similar.

The online classification was evaluated with the two best feature sets, stacked formants and stacked MFCCs, which achieve accuracy values of 84% and 79%, respectively.

5 Discussion

In this paper, we described a system for non-lexical confirmation detection in speech. Our system is capable of both online and offline processing of speech data. Thus, it can easily be integrated into systems interacting with humans. We relied on Support Vector Machines with a RBF kernel for classification. A sliding window approach enables the system to spot filled-pauses within utterances without the necessity to explicitly model parts of speech not relevant for filled-pause detection. The system's performance was evaluated on seven different feature sets: MFCCs, MFCCs with first and second derivative (Δ, $\Delta\Delta$), stacked MFCCs, formants, stacked formants, pitch and stacked pitch. The results show that successfully detecting non-lexical confirmations requires several frames of context. For this the stacking of the features improves the results and outperforms feature sets with derivatives. The results with stacked MFCCs, and MFCC related feature sets in general, are more stable within several performed test runs. But stacked formants have the potential to achieve higher classification results depending on the data. The amount of available data for SVM training might also influence the performance of the stacked feature sets and has to be evaluated.

Our approach can be applied to spot other acoustic events in speech data. In further studies, we aim to apply stacked features for the detection of other non-lexical speech events such as filled-pauses and for detecting socio-emotional sig-

nals such as uncertainty. Virtual agents like "BILLIE" will become more and more natural interaction partners by integrating those cues.

Acknowledgments

The authors gratefully acknowledge the German Federal Ministry of Education and Research (BMBF) for providing funding to project KOMPASS (FKZ 16SV7271K), within the framework of which our research was able to take place. This work was supported by the Cluster of Excellence Cognitive Interaction Technology "CITEC" (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Furthermore, the authors would like to thank our student worker Kirsten Kästel for data annotation.

References

Andersen, G., and Thorstein, F. 2000. *Pragmatic Markers and Propositional Attitude*. Amsterdam: John Benjamins. chapter Introduction, 1–16.

Audhkhasi, K.; Kandhway, K.; Deshmukh, O. D.; and Verma, A. 2009. Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4857–4860.

Bell, L.; Boye, J.; and Gustafson, J. 2001. Real-time handling of fragmented utterances. In *Proc. NAACL workshop on adaptation in dialogue systems*, 2–8.

Blackman, R. B., and Tukey, J. W. 1959. *Particular Pairs of Windows*. Dover. 98–99.

Bogdanov, D.; Wack, N.; Gómez, E.; Gulati, S.; Herrera, P.; Mayor, O.; Roma, G.; Salamon, J.; Zapata, J.; and Serra, X. 2013. Essentia: An open-source library for sound and music analysis. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, 855–858. New York, NY, USA: ACM.

Brennan, S. E., and Williams, M. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers.

Brossier, P. M. 2007. *Automatic Annotation of Musical Audio for Interactive Applications*. Ph.D. Dissertation, Queen Mary, University of London.

Fetzer, A., and Fischer, K. 2007. *Lexical Markers of Common Grounds*. London: Elsevier. chapter Introduction, 12–13.

Garg, G., and Ward, N. 2006. Detecting Filled Pauses in Tutorial Dialogs. (0415150):1–9.

Gibbon, D., and Sassen, C. 1997. Prosody-particle pairs as dialogue control signs. In *Proc. Eurospeech*.

Goldwater, S.; Jurafsky, D.; and Manning, C. D. 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication* 52(3):181 – 200.

Goto, M.; Itou, K.; and Hayamizu, S. 1999. A real-time filled pause detection system for spontaneous speech recognition. *Proceedings of EUROSPEECH* 227–230.

Guennebaud, G.; Jacob, B.; et al. 2010. Eigen v3. http://eigen.tuxfamily.org.

Harris, F. J. 1978. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE* 66:51–83.

Heck, M.; Mohr, C.; Stüker, S.; Müller, M.; Kilgour, K.; Gehring, J.; Nguyen, Q. B.; Nguyen, V. H.; and Waibel, A. 2013. *Segmentation of Telephone Speech Based on Speech and Non-speech Models*. Cham: Springer International Publishing. 286–293.

Kelley, J. F. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.* 2(1):26–41.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10:1755–1758.

Pearson FRS, K. 1901. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2(11):559–572.

Philippsen, A., and Wrede, B. 2017. Towards Multimodal Perception and Semantic Understanding in a Developmental Model of Speech Acquisition. Presented at the 2nd Workshop on Language Learning at Intern. Conf. on Development and Learning (ICDL-Epirob) 2017.

Savitzky, A., and Golay, M. J. E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8):1627–1639.

Tsiaras, V.; Panagiotakis, C.; and Stylianou, Y. 2009. Video and audio based detection of filled hesitation pauses in classroom lectures. In *2009 17th European Signal Processing Conference*, 834–838.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Yaghoubzadeh, R.; Buschmeier, H.; and Kopp, S. 2015. Socially cooperative behavior for artificial companions for elderly and cognitively impaired people. In *Proceedings of the 1st International Symposium on Companion-Technology*, 15–19.