Application of Image Processing Methods to Filled Pauses Detection from Spontaneous Speech

Dmytro Prylipko, Olga Egorow, Ingo Siegert, Andreas Wendemuth

Cognitive Systems Group, Otto von Guericke University Magdeburg, Germany
dmytro.prylipko@ovgu.de

Abstract

To obtain a more human-like interaction with technical systems, those have to be adaptable to the users individual preferences, and current emotional state. In human-human interaction the behaviour of the speaker is characterised by semantic and prosodic cues, given (among other indicators) as short feedback signals. These so called filled pauses minimally convey certain dialogue functions such as attention, understanding, confirmation, or other attitudinal reactions. These signals play a valuable role in the progress and coordination of interaction. Hereby, the first step enabling an automatic system to react on these signals is the detection of them within the users utterances. This is a quite complex task, as the filled pauses are phonetically short, consisting mostly only of one vowel and one consonant. In this paper we present our methods to detect filled pauses in a naturalistic interaction utilising the LAST MINUTE corpus. We used an SVM classifier and improved the results further, by applying a Gaussian filter to infer temporal context information and performing a morphological opening to filter false alarms. We obtained recall of 70%, precision of 55%, and AUC of 0.94.
Index Terms: Fillers, filled pause detection, non-verbal vocalisations, morphological filters, erosion, dilation

1. Introduction

Spoken communication of humans consists of several information layers, revealing details beyond the pure textual information given. These details are provided by conversationalists to enhance the interaction. Hereby, several semantic and prosodic cues are exchanged among the interaction partners and used to signalise the progress of the dialogue [1]. Among others, non-verbal vocalisations like filled pauses enrich the spoken words content with paralinguistic information, which is vital for determining speaker's state and intention underlying the utterance [2]. In order to enhance the human-computer interaction (HCI), an adaptation to the user's behaviour and the integration of a general human behaviour understanding is indispensable [3]. HCI systems have failed to note and respond to these details so far, resulting in users trying to cope with and adapt to the machine's behaviour.

Filled pauses (non-verbal vocalisations like /uh/ or /uhm/) are among the most frequently occurring features of spoken language [4]. While they are not intentionally produced, fillers are an important part of human-human interaction (HHI) [5]. The two most important functions are the communicative functions (like keeping the conversational turn while thinking) and affective functions (like expressing mental states) [6]. Batliner et al. [7] state that fillers indicate planning, turn taking, etc. and are cue phrases for repetitions and repairs. Swerts found that they highlight discourse structure and occur at major discourse boundaries [8]. According to Nicholson et al. [9], fillers also fulfil an interpersonal role. Furthermore, as filled pauses can occur at every point of spontaneous speech, they have a remarkable influence on the performance of speech recognition systems for spontaneous conversations.

During previous years, several methods for filled pauses detection have been presented. In [10] the authors presented a filled pause detection system based on combination of F0, duration and spectral analyses. In [6], filled pauses are detected on a basis of two features (small fundamental frequency transition and small spectral envelope deformation). The SRI group proposed a detection method based on classification of word boundaries produced by a speech recogniser [11]. Stouten and Martens developed a detection system in order to improve the speech recogniser performance [12].

The INTERSPEECH 2013 Paralinguistic Challenge [13] aimed to bring the detection problem onto a new level by providing a standardised corpus and a reference system. The winners of the Social Signals Sub-Challenge introduced a system, built upon a DNN classifier complemented with time series smoothing and masking [14]. In this paper we present our approach to the filled pauses detection and localisation using an SVM classifier and two methods of signal processing to smooth the results: Gaussian filter and morphological opening.

2. LAST MINUTE corpus

In this study we utilise the LAST MINUTE corpus [15] of naturalistic recordings described in [16] [17], and [18], which contains multimodal recordings of 133 German speaking subjects in a so called Wizard-of-Oz (WoZ) experiment. The setup revolves around an imaginary journey to Waiuku. Each experiment takes about 30 min. Using voice commands, the subjects have to prepare the journey, equip the baggage, and select clothing. More details on the design of the corpus can be found in [15]. While recruiting the subjects, an equal distribution of age, gender, and educational level was aimed at. The subjects are divided in two sub-groups according to their ages. The young group ranges from 18-28 years, the elder group consists of subjects over 59 years.

From the 133 dialogues we chose 89 with the recordings available in good quality. From that selection we employed 86 dialogues that have filled pauses annotated. The total duration of the subset is about 25 hours. This data was then divided into three sets: a train set comprising 56 dialogues, a development, and a test set, consisting of 15 dialogues each. The age and gender distribution of the speakers over subsets is shown in Table 1.

Table 1: *Distribution of subjects among age/gender groups.*

	train	dev	test	Total
Male	26	7	7	40
– young	*15*	*3*	*3*	*21*
– elderly	*11*	*4*	*4*	*19*
Female	30	8	8	46
– young	*13*	*4*	*5*	*22*
– elderly	*17*	*4*	*3*	*24*
All	56	15	15	86

The data has been separated into two classes, namely "Filler" and "Other". While the former consists of the filled pauses data only, the latter comprises the rest of the frames, including silent pauses. The sets contains an almost equal percentage of filled pauses: 7.35% of the time on the train set, 8.07% on the development set, and 8.02% on the test set (cf. Table 2).

Table 2: *Distribution of classes 'Filler' and 'Other' among training, development, and test sets.*

	Train	Development	Test	Total
		Time, s		
Filler	441	123	122	686
Other	5 181	1 396	1 402	7 979
		Frames		
Filler	44 128	12 247	12 296	68 671
Other	518 122	139 695	140 247	798 064
		Percentage		
Filler	7.35%	8.07%	8.02 %	7.92%
Other	92.65%	91.93%	91.88%	92.08%

All the fillers were pre-annotated using forced alignment based on literal transcriptions. The annotation was then manually corrected. In total, the data set we used comprises 1314 filled pauses with a total duration of 656 seconds. The duration of a single filled pause lies between 0.13 s and 2.17 s, the average duration is 0.52 $\pm$ 0.25s. The occurrence of this kind of feedback signals and their different meanings within these corpora was presented in [19].

Since the classes were not balanced and the data contained about ten times more "Other" instances than "Filler" ones, we downsampled the train set to avoid the bias towards the class "Other". In order to obtain an almost fifty-fifty distribution of the classes, we created ten subsets of the train set, each containing randomly chosen 10% of the instances of the class "Other" and all the data of the class "Filler". Among the downsampled sets, 45.98% of instances from each subset belong to the "Filler" class and 54.01% belong to the class "Other": 44 128 and 51 841 instances, respectively. In the experimental part, we use the downsampled training sets to train the classifier. Measures on the test set are reported in terms of mean and standard deviation over the ten evaluations using classifiers trained on ten training subsets.

3. Fillers detection from speech

The baseline system we started with is based on the framewise classification using an support vector machine (SVM) classifier. Comparison to the multilayer perceptron (MLP) in our unreported study shows that the SVM provides better detection accuracy and less gap between the recall and precision. We chose the LibSVM implementation (cf. [20]) of an SVM classifier with RBF kernel. While standard SVM only predicts the target class, LibSVM provides us with the probability estimates calculated using the pairwise coupling (more details are in [21]). These estimates are further used for post-processing. The optimal values of γ and cost parameter C (γ = 0.05 and C = 5.0) for the classifier were determined using grid search on the development set.

The feature set is the one used for the INTERSPEECH 2013 Social Signals Sub-Challenge [13]. Features were extracted with the openSMILE toolkit [22] on the frame-level basis (25ms window, 10ms shift). This set is derived from 47 low-level descriptors (LLDs): 12 mel-frequency cepstral coefficients (MFCCs) and logarithmic energy are computed along with their first and second order delta and acceleration coefficients providing us with the 39 features typically used for speech recognition. They are complemented with voicing probability, harmonics-to-noise ratio, F0 and zero-crossing rate, together with their deltas. For each frame-wise LLD the arithmetic mean and standard deviation across the frame itself and eight of its neighbouring frames (four before and four after) are used as the actual features. This results in 47 $\times$ 3 = 141 values per frame.

3.1. Smoothing with Gaussian filter

From Fig. 1 one can see that the baseline system features a high false alarm rate, resulting in high recall and low precision.

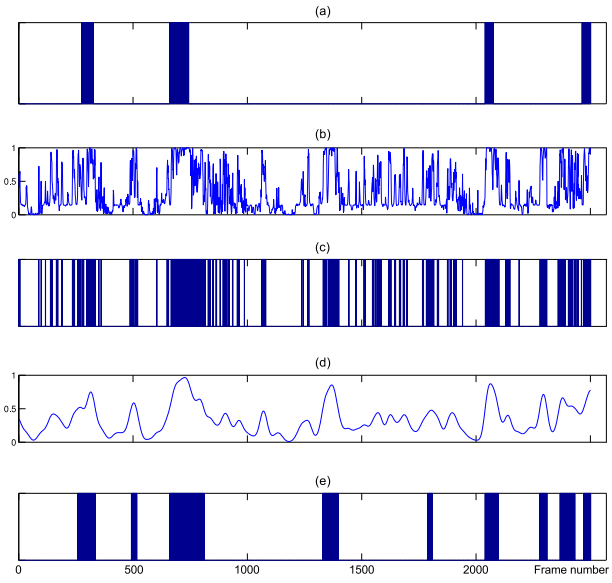

Figure 1: *From top to down: Reference (a), probability estimates from SVM (b), frame-wise localisation using SVM (c), probability estimates smoothed with Gaussian filter (d), localisation based on smoothed probability estimates (e). Solid bars correspond to fillers.*

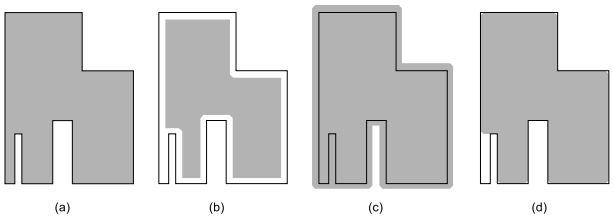

Figure 2: *From left to right: the original image X (a), its erosion (b), dilation (c), and opening (d) by a disk.*

In order to smooth the spikes and remove outliers on the probability estimates we used a Gaussian filter, widely employed for noise removal in signal and image processing. Filtering also brings some contextual information, which is lost in case of frame-wise classification. The parameters (window size and σ) of the Gaussian filter were tuned on the development set. Optimal values were found to be: window width of 40 frames and $\sigma = 10$.

After smoothing by the filter, the probability estimates lost the major erratic fluctuations, which resulted in less fragmented localisation (cf. Fig. 1). In turn, this improved both recall and precision.

3.2. Post-processing using morphological opening

The second stage of post-processing was performed by morphological opening – a method from mathematical morphology widely used in computer vision and image processing fields for noise removal [23, 24]. A convenient way to represent a binary (black and white) image is to define it as a subset of an Euclidian space $E = \mathbb{Z}^2$. Such a definition is easily transferable to the event spotting domain due to the binary nature of the output. The key idea underlying mathematical morphology is to probe an image $X \subseteq E$ with another small, pre-defined set A, called *structuring element* (cf. [25]). By "probing" it is meant whether the set A_h hits X (i.e., $A_h \cap X \neq \varnothing$), misses X (i.e., $A_h \cap X = \varnothing$), or lies entirely inside X (i.e., $A_h \subseteq X$). Here A_h denotes the translate of A along the vector h $A_h = \{a + h | a \in A\}$. Widely used structuring elements are disks, squares, crosses, and other geometrical primitives of different sizes.

In mathematical morphology, opening is a combination of two morphological operations: erosion and dilation. The *erosion* of the binary image X by the structuring element A is defined by:

$$X \ominus A = \{h \in E | A_h \subseteq X\} \tag{1}$$

The erosion process allows to make contours skinnier and to remove spikes and outliers (see Fig. 2). *Dilation* of the image X by the structuring element A is defined by:

$$X \oplus A = \bigcup_{a \in A} X_a \tag{2}$$

Dilation is the dual operation of the erosion. It can be used for smoothing and stacking of separate image elements. Finally, *opening* is the dilation of the erosion of the image X by the structuring element A:

$$X \circ A = (X \ominus A) \oplus A \tag{3}$$

For morphological opening we tested different kinds of structuring elements. The choice of the structuring element influences recall and precision in a certain way. If the aim is to increase the recall, then the opening should be done with a small structuring element or should not be done at all, since it will erase small pieces that overlap with the filled pauses. On the other hand, a larger structuring element will remove more false alarms, resulting in higher precision.

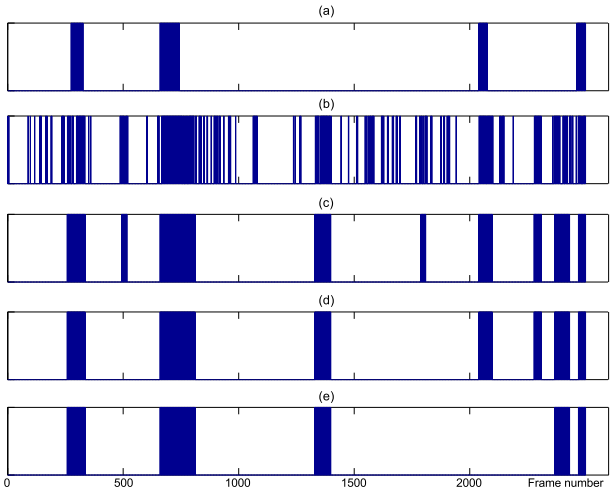

Figure 3: *Different stages of localisation. From top to down: Reference (a), frame-wise localisation by SVM (b), localisation using probability estimates smoothed by Gaussian filter (c), localisation post-processed with opening using small structuring element S_{22} (d), localisation post-processed with opening using larger structuring element S_{27} (e).*

For our task we found the square structuring element to provide the best F-score on the development set. Before finding the optimal structuring element we also tested other shapes (diamond and rectangle) and sizes on the development set. In the end, we employed square elements of two different sizes: The size of 22 frames was found to provide the optimal result, so the first structuring element is the square of size 22 (S_{22}). From the average duration of a filled pause (0.52s = 52 frames) and its standard deviation (0.25s = 25 frames) we can conclude that chunks of $52 - 25 = 27$ frames and less are likely to be false alarm. Thus, the second structuring element has size of 27 frames (S_{27}). Fig. 3 illustrates the application of morphological opening using the two structuring elements to the localisation results. It can be seen that employment of a larger structuring element caused removal of a correctly detected filler, while a false alarm has also been erased.

4. Results and discussion

As expected, the baseline system provided a very fragmented result with a high false alarm rate but also a high detection rate (see Table 3). Therefore our main goal was to reduce the false alarm and to increase the precision without losing the high recall. In Table 3 we can see that the application of the Gaussian filter allows us to achieve an improvement of 8.7% absolute for recall and of 10.3% for precision. The morphological opening did not provide such a substantial improvement, but it was shown to be useful for making the detection more balanced and to deliver a better F-score.

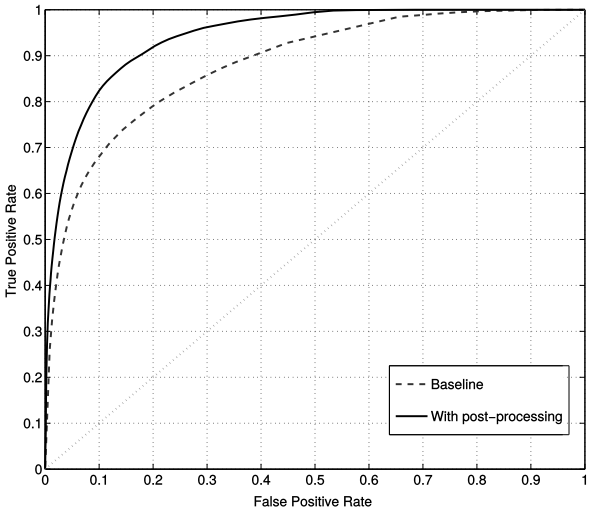

Figure 4: *Averaged ROC curves depending on the decision threshold ranging from 0 to 1. Post-processing is performed with Gaussian filter and structuring element S_{22}*

Varying the decision threshold on SVM output (default is 0.5), we can build a receiver operating characteristic (ROC) curves depicted on Fig. 4. The corresponding averaged areas under curve (AUC) are $0.88 \pm 7.4 \cdot 10^{-4}$ for the baseline system and $0.94 \pm 6.8 \cdot 10^{-4}$ for the classifier with post-processing. Equal error rates (EER) are 20.52% and 13.41% for the baseline and the final systems, respectively.

Table 3: *Detection accuracies of the evaluated system. Results for test set are given as the mean and standard deviation over ten evaluations using ten training subsets.*

	Recall, %	Precision, %	F-score, %
Development set			
Baseline	58.1	30.3	39.8
+ Gauss. filter	59.7	44.1	50.7
+ Opening			
– S_{22}	57.2	45.9	50.9
– S_{27}	54.4	46.6	50.2
Test set			
Baseline	66.1±0.2	39.6±0.3	49.6±0.3
+ Gauss. filter	74.8±0.3	49.9±0.4	59.8±0.3
+ Opening			
– S_{22}	72.6±0.3	52.6±0.5	61.0±0.3
– S_{27}	70.4±0.3	54.8±0.4	61.6±0.3

As it was discussed in Section 3.2, by choosing among the sizes of the structuring element, we can shift our system towards either higher recall or better precision. Such a precision-recall trade-off is illustrated in Fig. 5.

5. Conclusion

In this paper, we evaluated the system for filled pause detection and localisation from spontaneous speech. The difficulty of detecting fillers lies in their nature. For instance, they have the same phonetic characteristics as the frequently occurring phoneme `E:`, which makes the classification difficult and causes

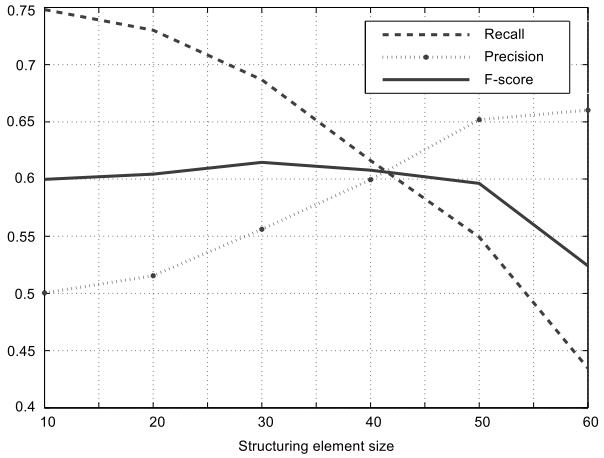

Figure 5: *Recall, precision, and F-score for class "Filler" depending on size of the structuring element (square).*

high false alarm rates for the baseline system. That is why we adhere to an opinion that usage of contextual information is indispensable for this task. We introduced the context with using the two techniques from signal and image processing, namely Gaussian filtering and morphological opening. Both methods were applied to post-process the probability estimates from the baseline system, built on an SVM classifier. Smoothing the class probability estimates with a Gaussian filter provided us with the most substantial improvement. The second stage of post-processing – morphological opening – allowed us to reduce the false alarm rate and to improve the detection precision and F-score. Also, morphological opening was found to be flexible tool for trade-off between the recall and precision.

Although the results are promising, there is still a room for further improvement, because the false alarm remains a problem, causing low precision score. We are aiming to address this issue with the introduction of more sophisticated contextual information. We believe that pitch contour analysis could improve the prediction of the fillers location and boundaries. Also, usage of linguistic information can be useful in separation between the standalone non-verbal vocalisations and the vowels within words, thus making the detection and localisation of fillers more accurate.

6. Acknowledgements

The work presented in this paper was done within the Transregional Collaborative Research Centre SFB/TRR 62 'Companion-Technology for Cognitive Technical Systems' funded by the German Research Foundation (DFG).

7. References

[1] J. Allwood, J. Nivre, and E. Ahlsen, "On the Semantics and Pragmatics of Linguistic Feedback," *Journal of Semantics*, vol. 9, no. 1, pp. 1–26, 1992.

[2] N. Campbell, "On the Use of NonVerbal Speech Sounds in Human Communication," in *Verbal and Nonverbal Communication Behaviours*, ser. Lecture Notes in Computer Science, A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro, Eds. Berlin, Heidelberg: Springer, 2007, vol. 4775, pp. 117–128.

[3] A. Wendemuth and S. Biundo, "A Companion Technology for Cognitive Technical Systems," in *Cognitive Behavioural Systems*, ser. Lecture Notes in Computer Science, A. Esposito, A. M.

Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds. Berlin, Heidelberg: Springer, 2012, vol. 7403, pp. 89–103.

[4] R. L. Rose, "The communicative value of filled pauses in spontaneous speech," Ph.D. dissertation, the University of Birmingham, 1998.

[5] M. Corley and O. W. Stewart, "Hesitation Disfluencies in Spontaneous Speech: The Meaning of um," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

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

[7] A. Batliner, A. Kiessling, S. Burger, and E. Nöth, "Filled pauses in spontaneous speech," in *Proceedings of ICPhS*, vol. 3, 1995, pp. 472–475.

[8] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, no. 4, pp. 485–496, Oct. 1998.

[9] H. Nicholson, K. Eberhard, and M. Scheutz, "Um... I dont see any: The Function of Filled Pauses and Repairs," in *Proceedings of the DiSS-LPSS Joint Workshop 2010*, Tokyo, Japan, 2010, pp. 89–92.

[10] M. Gabrea and D. O'Shaughnessy, "Detection of filled pauses in spontaneous conversational speech," in *Proceedings of ICSLP'2000*, vol. 3, 2000, pp. 678–681.

[11] E. Shriberg, R. A. Bates, and A. Stolcke, "A prosody only decision-tree model for disfluency detection," in *Proceedings of EUROSPEECH'97*, vol. 5, 1997, pp. 2383–2386.

[12] F. Stouten and J.-P. Martens, "A feature-based filled pause detection system for Dutch," in *Proceedings of ASRU'03*. IEEE, 2003, pp. 309–314.

[13] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, and Others, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH*, 2013.

[14] R. Gupta, K. Audhkhasi, S. Lee, and S. Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking," in *Proceedings of INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 173–177.

[15] D. Rösner, R. Friesen, M. Otto, J. Lange, M. Haase, and J. Frommer, "Intentionality in interacting with companion systems an empirical approach," in *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, ser. LNCS. Berlin, Heidelberg: Springer, 2011, vol. 6763, pp. 593–602.

[16] J. Frommer, B. Michaelis, D. Rösner, A. Wendemuth, R. Friesen, M. Haase, M. Kunze, R. Andrich, J. Lange, A. Panning, and I. Siegert, "Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus," in *Proceedings of LREC'12*, N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: ELRA, 2012, pp. 3064–3069.

[17] D. Rösner, J. Frommer, R. Friesen, M. Haase, J. Lange, and M. Otto, "LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions," in *Proceedings of LREC'12*, N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: ELRA, 2012, pp. 2559–2566.

[18] D. Prylipko, D. Rösner, I. Siegert, S. Günther, R. Friesen, M. Haase, B. Vlasenko, and A. Wendemuth, "Analysis of significant dialog events in realistic human-computer interaction," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 75–86, 2014.

[19] I. Siegert, K. Hartmann, D. Philippou-Hübner, and A. Wendemuth, "Human Behaviour in HCI: Complex Emotion Detection through Sparse Speech Features," in *Human Behavior Understanding*, ser. Lecture Notes in Computer Science, A. Salah, H. Hung, O. Aran, and H. Gunes, Eds. Springer International Publishing, 2013, vol. 8212, pp. 246–257.

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

[21] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. of the international conference on Multimedia (MM'10)*. ACM, 2010, pp. 1459–1462.

[23] J. Serra, "Introduction to mathematical morphology," *Computer vision, graphics, and image processing*, vol. 35, no. 3, pp. 283–305, 1986.

[24] J. Gil and R. Kimmel, "Efficient dilation, erosion, opening, and closing algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1606–1617, 2002.

[25] H. J. A. M. Heijmans, "Mathematical morphology: a modern approach in image processing based on algebra and geometry," *SIAM Review*, vol. 37, no. 1, pp. 1–36, 1995.