DNN-based Feature Extraction for Conflict Intensity Estimation from Speech

Gábor Gosztolya, and László Tóth, *Member, IEEE*

***Abstract*—Over the past few years there has been an increasing need to extract non-linguistic information from audio sources. This trend has created a new area in speech technology known as computational paralinguistics. A task belonging to this area is to estimate the intensity of conflicts arising in speech recordings, based only on the audio information. It was shown that the human comprehension of conflict intensity is closely related to speaker overlap; that is, when multiple persons are speaking at the same time. This type of information can also aid automated conflict intensity estimation. In this study we propose a simple, DNN-based feature extraction step, and show that this approach is superior to those introduced in the literature so far: by combining our results with an efficient greedy feature selection algorithm, we were able to outperform all previous results on the SSPNet Conflict dataset, achieving a correlation coefficient of 0.856 on the test set.**

***Index Terms*—computational paralinguistics, conflict intensity estimation, Deep Neural Networks, feature extraction**

I. Introduction

WITHIN speech technology, an emerging area is computational paralingustics, which seeks to detect, extract and locate non-linguistic information from the speech signal. Notable examples for paralinguistic tasks are emotion detection [1], detecting vocalizations such as laughter and filler events [2], [3], [4], and various medical applications like detecting Parkinson's or Alzheimer's disease or depression [5], [6], [7].

A specific paralinguistic task is to estimate the level of conflict from speech. Conflicts influence the everyday lives of people to a significant extent, either in their public or personal lives, and they are one of the main causes of stress [8]. With the rise of socially intelligent technologies, the automatic detection of conflicts could be the first step towards handling them properly. Furthermore, conflict detection has straightforward applications such as monitoring incoming calls in call centres, where a key feedback of the employees is how they can handle conflicted situations [9].

The standard computational approach, developed over the years on various paralinguistic tasks, is to extract several thousand general, utterance-level features from the speech excerpts, and use these to train general machine learning methods such as Support-Vector Machines (SVM) or Deep Neural Networks (DNNs) to perform classification or regression. Usually, however, other task-specific steps are required to achieve state-of-the-art performance, like feature selection (e.g. [10], [11]), incorporating features which are widely-used in other audio-based areas (e.g. i-vectors [12]), or developing new features for the given specific task (e.g. [13], [14]).

For conflict detection, a specific phenomenon which might aid detection is speaker overlap: in the heat of the debate, people tend to interrupt each other quite frequently, and speak while someone else is speaking. There were several studies which exploited this observation for conflict intensity estimation: Grèzes et al. [15] included the ratio of speaker overlap as a new feature in the baseline feature set. Brueckner and Schuller [16] used Deep Bidirectional Recurrent Neural Networks to estimate speaker overlap and used it as a feature along with other prosodic attributes; Caraty and Montacié [17] detected speech interruptions to aid the detection of utterances with a high level of conflict.

However, in our opinion these methods can only be applied in a limited way. Grèzes estimated the amount of speaker overlap by a simple procedure; using a BLSTM like Brueckner et al. may be viewed as an overkill for detecting speaker overlap due to implementation difficulties, while the bidirectional nature of his approach makes it unsuitable for real-time speech processing; the workflow proposed by Caraty and Montacié inherently works only for conflict *classification*, and does not allow finer intensity distinctions. In this study we propose a simple-yet-efficient approach, where neural networks are trained to detect local speaker overlap; then, for the next step, several features are extracted from the outputs of the DNN. We show that this approach leads to a better performance than using either the manually annotated or the predicted (single) speaker overlap values: by combining these predictions with those obtained by our feature selection method introduced earlier [18], we markedly outperform all previous results on a public database containing political debates.

II. The SSPNet Conflict Corpus

The SSPNet Conflict Corpus [19] contains recordings of Swiss French political debates taken from the TV channel "Canal9". It consists of 1430 recordings, 30 seconds each, making a total of 11 hours and 55 minutes. Each 30-second long clip was tagged by 10 annotators; in the end each recording was assigned a score in the range [-10, 10], 10 meaning a high level of conflict and -10 meaning no conflict at all. Although the database contains both audio and video recordings, in the recent experiments researchers focused only on the audio information. To demonstrate the effectiveness of our automatic speaker overlap detection method, here we will also rely on the audio data, and discard the video track.

G. Gosztolya was with the Institute of Informatics, University of Szeged, Hungary, e-mail: ggabor@inf.u-szeged.hu.

G. Gosztolya and L. Tóth were with the MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary.

Manuscript received April 19, 2005; revised August 26, 2015.

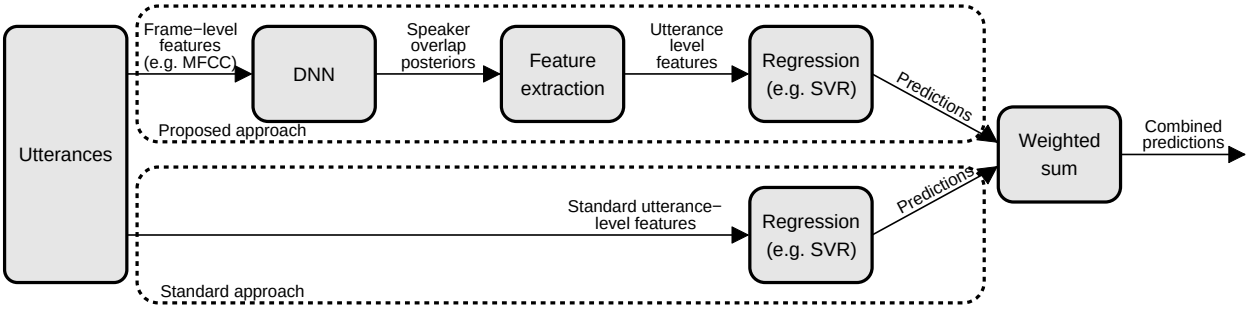

Fig. 1. The proposed regression approach (upper path), the standard paralingustic process (lower path), and the calculation of the final predictions.

The audio clips of this dataset were then used in the Conflict sub-challenge of the Interspeech 2013 ComParE Challenge [20]. Besides completely discarding video data, other steps were made to standardize the work on this dataset, and this setup has since been adopted by most researchers. Firstly, separate training and test sets were defined instead of relying on cross-validation, as was done by Kim et al. [19]; secondly, a baseline feature set was defined and extracted from the utterances by the tool openSMILE [21]. This 6373-long feature set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs), over which statistical functions (e.g. mean, standard deviation, etc.) are computed to provide utterance-level feature values.

The evaluation metrics used for this dataset were also defined. Schuller et al. admitted that this was mainly a regression task and used the correlation coefficient (CC) to measure the performance. They, however, also converted the task into a binary classification one, defining the classes *low* and *high* based on the sign of the conflict score [20]. Classification accuracy was measured by the Unweighted Average Recall (UAR) value; this metric was used both in the Challenge (e.g. [10], [15]), and it has been used in research papers since then (e.g. [11], [16], [17]). In our view, treating this task as a regression one is the proper approach, partly since describing conflict intensity as a numeric value contains more information than a binary class label, and also because optimizing for CC led to more robust models than maximizing UAR. (For the details, see [18].) Due to this, now we will use the CC metric.

Table I lists the notable scores published in the literature for this dataset. We can see two trends: most attempts either applied feature selection ([10], [11], [18]) or utilized the amount of speaker overlap in some way ([15], [16], [17]). Next we will propose a speaker overlap-based feature extraction step, and combine this approach with our previous one [18] where we used feature selection.

TABLE I
CORRELATION COEFFICIENT (CC) AND UAR SCORES GIVEN IN THE LITERATURE FOR THE TEST SET OF THE SSPNET CONFLICT CORPUS, FOLLOWING THE COMPARE 2013 SETUP. HERE, "—" MEANS THAT THE GIVEN SCORE WAS NOT PROVIDED.

Method	CC	UAR
ComParE 2013 baseline ([20])	0.816	80.8%
Speaker overlap (Grèzes, [15])	—	83.1%
Random Subset FS (Räsänen, [10])	0.826	83.9%
Speaker overlap + prosodic feat. (Brueckner, [16])	0.838	84.3%
SLCCA FS (Kaya, [11])	—	84.6%
Speaker Interruption (Caraty, [17])	—	85.3%
Greedy Forward FS (Gosztolya, [18])	0.835	85.6%
Greedy Forward + Backward FS (Gosztolya, [18])	0.842	85.1%
Ensemble Nyström method (Huang, [22])	0.849	—

III. SPEAKER OVERLAP-BASED FEATURE EXTRACTION

A high level of conflict frequently coincides with multiple persons speaking at the same time. Grèzes et al. demonstrated experimentally that exploiting the speaker overlap could aid the automatic estimation of conflict intensity: by extending the baseline ComParE feature set with the (predicted) relative amount of speaker overlap, they markedly outperformed the baseline scores [15]. Indeed, on this corpus we measured a correlation coefficient of 0.70 between the conflict score of the utterance and the relative amount of time when multiple speakers spoke at the same time (according to the manual annotation), which implies a very close connection. Of course, relying on a *manually annotated* speaker overlap value is not an option in an application situation. If we seek to utilize the amount of speaker overlap in the conflict intensity estimation task, we should calculate it in some automatic way.

Many studies exist which deal with automatic speaker clustering and diarization (e.g. [23], [24], [25], [26]); these, however, focus on finding the time intervals where the same speakers' voice is present, which is clearly not our main focus here. One can also argue that the amount of speaker overlap is reflected in the volume of the utterance: when two or more people are speaking at the same time, their (combined) volume can be expected to exceed that of only one speaker, and this local energy can be readily determined by signal processing techniques. Another option might be the one proposed by Grèzes et al. [15], who estimated speaker overlap from the utterance-level, 6373-sized feature set via standard regression.

In this study we propose another approach; for the general scheme of the proposed workflow, see Fig. 1. As the first step, we train a DNN to predict the number of actual speakers for each given frame. Then, in the second step, we extract a number of (utterance-level) features from the DNN outputs, which are used to train a Support-Vector Regression (SVR [27]) to predict the conflict intensity scores of the utterances. Lastly, we combine the predictions with the ones obtained using standard utterance-level features. We will see that this approach

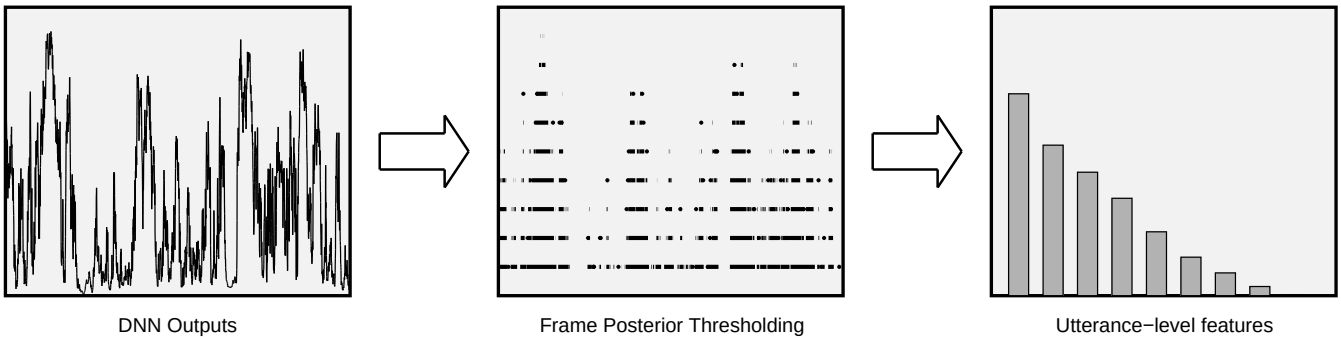

Fig. 2. The mechanism of the proposed feature extraction process.

outperforms all previous results on this public dataset. We will also show that the proposed approach is superior to using either speaker overlap as an additional feature, or to extracting utterance-level features from the energy of the utterance. For comparison, we also test the automatic estimation procedure proposed by Grèzes et al. [15].

A. DNN-Based Speaker Overlap Estimation

The first step of our proposed workflow is quite straightforward: we train a DNN with standard frame-level features (e.g. MFCC [28]) as input, while the output neurons correspond to the number of speakers in the given frame. In spontaneous speech it is quite rare that three or more people are speaking at the same time: according to the manual diarization, it does not happen in the SSPNet Conflict corpus at all. Due to this, we propose to use only two classes, corresponding to a zero-or-one speaker, and a two-or-more speaker case.

B. DNN-Based Feature Extraction and Regression

Next, we extract features from the frame-level DNN outputs, and these features will be used for *utterance-level* regression. Naturally, in general applications we should rather try to perform this regression step for a specific time window instead of the whole utterance. In the actual dataset, however, the manual annotation of the level of conflicts is given at the utterance level only, which does not permit continuous conflict intensity evaluation. However, our approach can be easily generalized into longer utterances by using sliding windows.

In the actual feature extraction phase, we seek to include the amount of time where two people were speaking at the same time. The most straightforward solution is to classify each frame based on the DNN outputs, and count the ratio of the frames classified as having multiple speakers present. Since we have two classes, this is equivalent to thresholding the corresponding DNN outputs with the value of 0.5 [29]. However, it is well known (see e.g. [30], [31]) that the posterior estimates provided by a DNN carry valuable information, and this information is lost if we simply examine whether they exceed 0.5. Because of this, we propose to use *several* different threshold values. That is, using the step size parameter s, first we count the number of frames where the DNN output corresponding to the two or more speakers case is greater than or equal to s, we divide it by the total number of frames in the utterance, and then use this value as the first newly extracted feature. Next, we repeat this step using the values $2 \cdot s, 3 \cdot s, \ldots, 1$ as thresholds. Doing this for all the utterances, we extract a new feature set for all the examples. This can be used as a feature set to perform regression for the third step by using some machine learning method like DNNs or SVR.

C. Regression Output Combination

Although using the amount of speaker overlap may prove to be beneficial for conflict intensity estimation, we should not discard all other kinds of features. A combination of the two approaches supposedly leads to better results. One possible way of combining them is to merge the *feature vectors* of each example, and train one classifier or regressor model. However, often (e.g. [32]) it is more beneficial to train separate machine learning models for different types of features, as these may require different meta-parameter settings for optimal performance. Therefore we suggest training one machine learning method using the standard utterance-level features such as that proposed in [20], and train a separate one using the features extracted as described in Section III-B. To combine *the outputs* of the two models, we suggest taking the weighted mean, which is a simple-yet-robust technique (see e.g. [32]).

IV. Experimental Setup

A. DNN Parameters

To predict the amount of speaker overlap, we trained a DNN with 5 hidden layers, each containing 256 rectified neurons [33]. We utilized our custom implementation for Nvidia GPUs; we used 39 MFCC + Δ + $\Delta\Delta$ [28] values as feature vectors on a 15-frame wide sliding window.

B. Feature Extraction and Regression

For the next feature extraction step, we used a step size of 0.05 for the thresholds, resulting in 20 features overall. After standardization (i.e. transforming the feature vectors so as to have a zero mean and unit variance), we trained a Support-Vector Regressor using the LibSVM [34] library. We applied the nu-SVR method with linear kernel; the value of C was tested in the range $10^{\{-5,\ldots,1\}}$, just like that in our previous paralinguistic studies (e.g. [18], [32], [35]). As the energy of the speech signal might also be an indicator of conflict, we performed the same thresholding feature extraction steps on the frame-level energy values to get 20 features overall.

C. Standard Paralinguistic Approach

We performed an estimation of conflict intensity scores following the standard paralinguistic approach as well. For this, we commenced with the default, 6373-long feature set proposed by Schuller et al. [20]. This feature set has a lot of redundancy, and also contains many irrelevant features for this given task. Owing to this, we decided not to use the full feature set for training the SVR, but used the restricted feature set chosen by the method proposed by Gosztolya [18]. This greedy forward-backward feature selection algorithm first sorts the features based on the absolute value of their correlation coefficient with the target score in descending order; this way it examines more correlated features first. Then it examines each feature in this order, and decides whether this particular feature should be selected or discarded based on whether it improved the regression performance of SVR on the development set. Next, a backward step is performed to prune this feature set further. (For the details, see [18].) The resulting feature set consisted of only 137 attributes out of the original 6373.

D. Prediction Combination

To combine the utility of different feature sets, we opted for two approaches. In the first one, we merged the feature sets, and trained only one SVR model for the combined feature set. In the second one we trained three SVR models using the three kinds of feature sets tested (i.e. the one got by feature selection, and the two sets extracted following Section III-B), and combined the predictions via a weighted mean. We determined the weights by grid search, using a step size of 0.05; we chose the weight vector that proved to be the best on the training set, using 10-fold cross-validation (CV).

V. Results

Table II shows the correlation coefficient values obtained in the cross-validation setup and on the test set. We see that by using the selected feature subset determined by the greedy feature selection method, we can markedly outperform the baseline score on the test set. (The indicated value is slightly lower than the one published in [18] because now we used ten-fold cross-validation instead of the development set.) Surprisingly, using only the features extracted from the DNN posteriors, we can almost match the baseline score: the 0.809 correlation coefficient measured on the test set is only slightly lower than the baseline value of 0.816. However, when we relied only on the energy of the speech signal, the results were much lower than those got by using the other approaches.

When we extended the standard features with the newly extracted ones, the CC values rose further. The speaker overlap ratio estimated from the 6373 utterance-level features ("predicted speaker overlap") did not help much when we used the selected feature subset, and it did not affect conflict intensity estimation performance when combined with the full feature set at all. Using the energy-based features led to similarly small improvements. However, the automatically extracted, DNN-based speaker overlap feature set helped as much as the manually annotated speaker overlap value did. On top of these, using the energy-based attributes did not really help.

TABLE II
Correlation coefficients obtained by the approaches tested.

Combination	Feature Set	CV	Test
—	Full feature set (baseline) [20]	0.830	0.816
—	Selected feature subset [18]	0.825	0.838
	Speaker overlap	0.771	0.809
	Energy	0.597	0.548
Feature set	Full + Predicted sp. overlap	0.830	0.816
	Selected + Manual sp. overlap	0.837	0.846
	Selected + Predicted sp. overlap	0.827	0.840
	Selected + Sp. overlap	0.837	0.846
	Selected + Energy	0.825	0.840
	Selected + Sp. overlap + Energy	0.838	0.846
Prediction	Selected + Sp. overlap	0.837	0.856
	Selected + Energy	0.826	0.837
	Selected + Sp. overlap + Energy	0.837	0.855

We got the best results when we trained separate SVR models for the different kinds of feature sets, and combined the outputs instead. The energy-related features were again of little use, but using the automatically determined speaker overlap scores was a big help: the 0.856 correlation coefficient obtained in this way on the test set is the highest such score ever published on this dataset. In our opinion it is due to the fact that we extracted a whole feature *set* describing speaker overlap instead of one single ratio value, and it contained more information. The optimal weight for the predictions by the proposed method was 0.3, showing that the feature selection approach is more important (its weight being 0.7), but the speaker overlap was also essential for state-of-the-art performance. This finding is also in accordance with the correlation scores got by using the two methods independently. These predictions had an UAR score of 84.7% on the test set, which, given that we optimized all meta-parameters for CC, is quite competitive. We would also like to note that the UAR scores varied to a significant extent, which is probably due to a number of predictions being close to zero, where their sign (and therefore their binary class label) can change easily; this, in our opinion, also supports our decision of utilizing CC instead of the UAR metric in this particular task.

VI. Conclusions

In computational paralinguistic tasks we need to perform task-dependent steps to achieve state-of-the-art accuracy scores. One such step that could aid conflict intensity estimation from the audio data is to estimate the duration of when two or more speakers were speaking at the same time. For this, we proposed a simple DNN-based feature extraction step, which not only returned the single value of the speaker overlap estimated, but also a 20-sized vector which characterizes speaker overlap in the utterance in a more sophisticated way. By using this novel feature extraction step, followed by a regression step and combining the prediction scores using a weighted mean, we achieved a marked improvement in our correlation coefficient scores: our 0.856 score is the highest one published so far on the public SSPNet Conflict corpus.

REFERENCES

[1] S. L. Tóth, D. Sztahó, and K. Vicsi, "Speech emotion perception by human and machine," in *Proceedings of COST Action*, Patras, Greece, 2012, pp. 213–224.

[2] T. Neuberger, A. Beke, and M. Gósy, "Acoustic analysis and automatic detection of laughter in Hungarian spontaneous speech," in *Proceedings of ISSP*, 2014, pp. 281–284.

[3] R. Brueckner and B. Schuller, "Social signal classification using deep BLSTM recurrent neural networks," in *Proceedings of ICASSP*, 2014, pp. 4856–4860.

[4] R. Gupta, K. Audhkhasi, S. Lee, and S. S. Narayanan, "Detecting paralinguistic events in audio stream using context in features and probabilistic decisions," *Computer, Speech and Language*, vol. 36, no. 1, pp. 72–92, 2016.

[5] I. Hoffmann, D. Németh, C. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in Alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[6] J.-R. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, and E. Nöth, "Analysis of speech from people with Parkinson's disease through nonlinear dynamics," in *Proceedings of NoLISP*, 2013, pp. 112–119.

[7] G. Kiss, M. G. Tulics, D. Sztahó, and K. Vicsi, "Language independent detection possibilities of depression by speech," in *Proceedings of NoLISP*, 2016, pp. 103–114.

[8] P. Spector and S. Jex, "Development of four self-report measures of job stressors and strain: interpersonal conflict at work scale, organizational constraints scale, quantitative workload inventory, and physical symptoms inventory," *Journal of Occupational Health Psychology*, vol. 3, no. 4, pp. 356–367, 1998.

[9] M. Koutsombogera, D. Galanis, M. T. Riviello, N. Tseres, S. Karabetsos, A. Esposito, and H. Papageorgiou, *Conflict Cues in Call Center Interactions*. Springer International Publishing, 2015, ch. 18, pp. 431–447.

[10] O. Räsänen and J. Pohjalainen, "Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech," in *Proceedings of Interspeech*, Lyon, France, Sep 2013, pp. 210–214.

[11] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Random discriminative projection based feature selection with application to conflict recognition," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, 2015.

[12] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 1402–1406.

[13] D. Sztahó, G. Kiss, and K. Vicsi, "Estimating the severity of Parkinson's disease from speech using linear regression and database partitioning," in *Proceedings of Interspeech*, Dresden, Germany, 2015, pp. 498–502.

[14] C. Montacié and M.-J. Caraty, "Prosodic cues and answer type detection for the deception sub-challenge," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2016–2020.

[15] F. Grèzes, J. Richards, and A. Rosenberg, "Let me finish: Automatic conflict detection using speaker overlap," in *Proceedings of Interspeech*, 2013, pp. 200–204.

[16] R. Brueckner and B. Schuller, *Be at Odds? Deep and Hierarchical Neural Networks for Classification and Regression of Conflict in Speech*. Springer International Publishing, 2015, ch. 19, pp. 403–429.

[17] M.-J. Caraty and C. Montacié, *Detecting Speech Interruptions for Automatic Conflict Detection*. Springer International Publishing, 2015, ch. 18, pp. 377–401.

[18] G. Gosztolya, "Conflict intensity estimation from speech using greedy forward-backward feature selection," in *Proceedings of Interspeech*, Dresden, Germany, Sep 2015, pp. 1339–1343.

[19] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli, "Predicting continuous conflict perception with Bayesian Gaussian Processes," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 187–200, 2014.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion, Autism," in *Proceedings of Interspeech*, 2013.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia*, 2010, pp. 1459–1462.

[22] D.-Y. Huang, H. Li, and M. Dong, "Ensemble nyström method for predicting conflict level from speech," in *Proceedings of APSIPA*, Siem Reap, city of Angkor Wat, Cambodia, Dec 2014, pp. 2418–2422.

[23] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of ASRU*, 2003, pp. 411–416.

[24] K. Yu, X. Jiang, and H. Bunke, "Partially supervised speaker clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 959–971, 2012.

[25] K. J. Han and S. S. Narayanan, "Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling," in *Proceedings of Interspeech*, 2008, pp. 20–23.

[26] A. Beke, "Automatic speaker diarization in Hungarian spontaneous conversations," Ph.D. dissertation, Eötvös Lóránd University, Budapest, Hungary, 2014.

[27] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[28] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[29] R. Busa-Fekete, B. Szörényi, K. Dembczyński, and E. Hüllermeier, "Online F-measure optimization," in *Proceedings of NIPS*, Cambridge, MA, USA, 2015, pp. 595–603.

[30] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[31] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, 2014.

[32] G. Gosztolya, T. Grósz, Gy. Szaszák, and L. Tóth, "Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2026–2030.

[33] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of AISTATS*, 2011, pp. 315–323.

[34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[35] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, "Determining native language and deception using phonetic features and classifier combination," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2418–2422.