*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Exploring Text and Audio Embeddings for Multi-Dimension Elderly Emotion Recognition

*Mariana Julião*[1,2], *Alberto Abad*[1,2], *Helena Moniz*[1,3]

[1]INESC-ID, Lisbon, Portugal
[2]Instituto Superior Técnico, University of Lisbon, Portugal
[3]FLUL - School of Arts and Humanities, University of Lisbon, Portugal
mariana.juliao@tecnico.ulisboa.pt, alberto.abad@inesc-id.pt, helena.moniz@inesc-id.pt

## Abstract

This paper investigates the use of audio and text embeddings for the classification of emotion dimensions within the scope of the Elderly Emotion Sub-Challenge of the INTERSPEECH 2020 Computational Paralinguistics Challenge. We explore speaker and time dependencies on the expression of emotions through the combination of well-known acoustic-prosodic features and speaker embeddings extracted for different time scales. We consider text information input through transformer language embeddings, both isolated and in combination with acoustic features. The combination of acoustic and text information is explored in early and late fusion schemes. Overall, early fusion of systems trained on top of hand-crafted acoustic-prosodic features (eGeMAPS and ComParE), acoustic model feature embeddings (x-vectors), and text feature embeddings provide the best classification results in development for both Arousal and Valence. The combination of modalities allows us to reach a multi-dimension emotion classification performance in the development challenge data set of up to 48.8% Unweighted Average Recall (UAR) and 61.0% UAR for Arousal and Valence, respectively. These results correspond to a 16.2% and a 8.7% relative UAR improvement.
**Index Terms**: computational paralinguistics, speech emotion recognition, speaker embeddings, text embeddings, elderly

## 1. Introduction

The Elderly Emotion Sub-Challenge of the INTERSPEECH 2020 Computational Paralinguistics Challenge (ComParE2020) [1] focuses on the emotion classification of spontaneous narratives uttered by German-speaking elderly. Two emotional dimensions are considered: Arousal – how affected the speaker is – and Valence – "positiveness" of the emotion [2], with levels mapped to Low, Medium, and High.

Speech Emotion Recognition and Understanding is a very active research field, in particular now that it is quite ubiquitous in people's lives. Virtual assistants or robots have a crucial role with elderly populations since they can serve as companions to assist the elder in several daily activities, towards healthy ageing and well-fare. However, assistants can only be seen as real companions if they understand the emotional states of their users and fine-tune their actions towards such states.

To address the challenges of emotion recognition proposed in the ComParE 2020 Elderly Emotion Sub-Challenge, we investigate the relevance and relations of audio and text features, speaker identity, time granularity and idiosyncratic traits of emotions. Our goal is to understand the impact of distinct features together with specific speaker-dependent characteristics on the recognition of Valence/Arousal. To this end, our proposed approach is structured as follows. First, we explore the acoustic feature sets. We investigate the performance of classic acoustic-prosodic feature sets for emotion classification and of acoustic speaker embeddings (x-vectors). As well, we explore the combinations of these two approaches. Then, we investigate the discriminative ability of long-context text embeddings based on transformers for the classification of emotions. Finally, we combine text and audio features to assess the possible complementary information in each of the targeted emotional dimensions. For that purpose, we consider two simple combination schemes based either on early fusion at the feature level or late fusion at the systems' decision level. Overall, the proposed system achieves up to 48.8% Unweighted Average Recall (UAR) and 61.0% UAR for Arousal and Valence multi-dimension emotion classification in the development challenge data set.

The rest of this document is organized as follows: Section 2 introduces the relevant state of the art. Then, in Section 3, we describe our methodology, including different types of features, fusion strategies, classifiers and additional insights on explored (unsuccessful) approaches. Experimental results are presented in Section 4 and the paper ends with conclusions in Section 5.

## 2. Related Work

The paper of Russel [3] presents a representation of emotional states as a two-dimensional space, each dimension corresponding to arousal and valence. Although criticized since then, namely for its incompleteness [4], it has been used in many emotion recognition tasks. For the task at hand, we are given a quantitative and dimensional representation of emotional states – i.e., emotions are represented as combinations of Arousal and Valence as Low, Medium and High.

The advent of Deep Neural Networks (DNN) brought a whole new insight on feature engineering. On the one hand, they have allowed for end-to-end classification approaches, with the possibility of feeding a model with raw input, without the need for precomputing features [5]. On the other hand, activations of the last layers of a DNN that has been tuned and trained for certain tasks have the potential to properly represent the classes being classified and, therefore, to be used as features (embeddings). For instance, the embeddings obtained from networks trained for speaker classification – x-vectors [6] – can be used as a sort of compact speaker representations. Furthermore, the use of x-vectors has been recently extended to other tasks, such as emotion classification [7] or pathological speech detection [8].

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), with reference UIDB/50021/2020, through PhD grant SFRH/BD/139473/2018.



http://dx.doi.org/10.21437/Interspeech.2020-2290

These neural representations have also played a remarkable role in Natural Language Processing tasks, as they have fuelled most of its recent achievements – from language modelling to machine translation [9]. Pre-trained models for ELMO [10], GPT [11] or BERT [12] allowed researchers to obtain better results with fewer data and less computation effort, as they can easily be fine-tuned and implemented for other tasks. Text embeddings have already been used for emotion classification, as in the case of [13].

Although audio and text can be used separately for emotion classification, it seems that these modalities do not have the same discrimination power for both dimensions of emotion. According to Karadogan and Larsen [14], valence is better recognized with semantic features, whereas arousal is better recognized with acoustic features. The authors explain it with the fact that valence is more related to the actual content of what is being said, while arousal is more associated with the way it is said. Ultimately, the combination of modalities was reported to provide better overall results than the ones provided by each modality alone.

Taking into account the positive impact that the use of feature embedding representations has recently shown in several tasks of speech and text classification, as well as the referred known complementary contribution of both types of modalities for emotion recognition, we focus here on the study of the possible synergies of this type of feature embeddings for the ComParE 2020 Elderly Emotion Sub-Challenge.

# 3. Methods

## 3.1. Acoustic features

All acoustic features are extracted for fixed-length audio chunks. The output of the classification of each narrative will then result from either voting the output of each chunk or from the classification of the averaged features of all chunks, as described next in Subsection 3.4.

### *3.1.1. Hand-crafted features*

In this work, we consider two sets of hand-crafted or knowledge-based engineered features extracted using the openSMILE toolkit [15]: ComParE2013 and eGeMAPS.

The ComParE Acoustic Feature Set [16], from the challenge of 2013, is a set of 6373 functionals over a set of low-level descriptors (LLD), which has mostly been used for paralinguistic tasks in the last decade. As a very complete set, many of its features can be redundant or irrelevant for certain tasks. As well, its size can sometimes be overwhelming for some classifiers.

The Geneva Minimalist Acoustic Parameter Set [17] is a small set of voice parameters, "based on a) their potential to index affective physiological changes in voice production, b) their proven value in former studies as well as their automatic extractability, and c) their theoretical significance" [17]. It comprises both the minimalist set of 62 features (GeMAPS) and the extended set of 88 features (eGeMAPS), the former appending spectral and frequency-related parameters to the minimalist set. Both have provided results close to those attained by ComParE feature set in previous benchmark comparisons [17]. Specifically, when averaging results over many databases [17], eGeMAPS reached the best Unweighted Average Recall (UAR) for Arousal (79.71%), whereas ComParE reached the best UAR for Valence (67.17%). In this work, we have used the set of 88 features, eGeMAPS.

### *3.1.2. Speaker embeddings: x-vectors*

X-vectors are fixed-dimensional embeddings, extracted from a deep neural network, which takes as input variable-length utterances and are trained to discriminate between speakers [6].

In this work, we extracted 512-dimensional x-vectors for the given audio chunks using a Kaldi speech recognition toolkit [18] model, pre-trained with the new Vox Celeb dataset (v2) [6, 19]. As the expression of emotion may be speaker-dependent, we decided to explore the potential of these specific speaker-modelling features to perform a sort of speaker-wise normalization of the acoustic feature vectors. Thus, we aim at exploring ways of combining previous hand-crafted features obtained for each chunk with speaker embedding information. To do so, we have first computed an average x-vector for each speaker over all their chunks. Then, we have explored the following three different ways of looking into the dependency of emotions on speaker and on time:

1. *on-line x-vector* (oXv) – x-vector obtained from the same audio chunk;
2. *normalized on-line x-vector* (nXv) – x-vector obtained from the same audio chunk normalized by the corresponding average speaker x-vector;
3. *speaker x-vector* (sXv) – average speaker x-vector.

In the first case, we expect the embeddings to capture the specific speaker-state information that is present in each short segment. In the second case, although the goal is the same, we attempt to emphasize the specific speaker state changes by removing the average speaker characteristics. In the later, by feeding the classifier with constant speaker side information, we expect it to be able to find common characteristics in the way groups of speakers produce their emotions, thus, attaining some sort of speaker adaptive training similar to what is done in automatic speech recognition [20]. In addition to the combination of speaker embeddings with hand-crafted features, in the experimental section we also investigate the potential of online x-vectors (oXv) and normalized on-line x-vectors (nXv) as individual feature extractors for emotion recognition.

## 3.2. Text Features

According to the recent research in pre-trained language models, multilingual models are able to capture inherent representations of language and, therefore, to support generalization to other languages [21]. For our problem, we foresaw that sources on which German monolingual models had been trained might not have a sufficient representation for the elderly population. Whereas the same is probably the case for multilingual models, as these have shown great ability to generalize, we decided to use the multilingual one. Indeed, to extract text features, we used bert-as-a-service [22], with a pre-trained Multilingual Cased model of 104 languages, 12 layers, 768 hidden states, 12 heads and 110M parameters.

## 3.3. Classification

For classification, we used the same approach as the official challenge baseline [1]. This consists of a linear SVM classifier, input data scaling to zero mean and unit standard deviation, class upsampling to natural factors in train, and complexity optimization.

All reported results correspond to the development partition, with classifiers trained on the train partition. We ran all classifiers for 6 different values of the complexity parameter:

$1^{-5}$, $1^{-4}$, $1^{-3}$, $1^{-2}$, $1^{-1}$, 1. Each of the results we present here corresponds to the best result out of all results obtained for the same features and different values of complexity.

### 3.4. Narrative predictions generation

In this task, each narrative is given one label for arousal and one label for valence (with levels high, medium or low). The audio of the initial narrative is released as chunks of 5 seconds, which are assigned to the same labels as the entire narrative. Consequently, since classification happens narrative-wise, it is necessary to convert the chunk predictions into a single global narrative prediction. To address this, the baseline uses Majority Voting, assigning to each narrative the most frequently assigned label to its chunks. The UAR is then computed overall narratives. Majority Voting is, thus, our first way of having one label per narrative.

Then, to allow for the combination of acoustic and text modalities at the feature-level (v. Subsection 4.4), all feature vectors for one narrative are summarized into one single feature vector by means of simple feature vector averaging. Thus, we took this as another way of having one single classification for one narrative: classifiers are trained and tested on single narrative averaged vectors.

### 3.5. Left-Behind Approaches

In the course of the current work, we explored some additional techniques which provided no remarkable results in terms of emotion classification, but which are, nevertheless, worth mentioning:

- *Statistical analysis of the text:* We explored 1) the correlation between labels and the rates of disfluencies (as "äh" or "ähm"), 2) the variation of the text-size for each speaker across emotional states, 3) the average size for each category, in general.
- *Time-informativeness:* With the intuition that not all parts of the same narration are equally informative regarding its emotional state, we investigated whether some parts (audio chunks) tended to be better predicted.
- *Summarization:* For audio chunks, we expected that the different text segments had a different contribution on the overall informativeness of the text. Thus, we attempted to create embeddings from a summarized version of the text obtained with the TF-IDF algorithm (hoping that fewer segments would reduce the disparity amongst them in terms of relevance).

Despite the discouraging results obtained with those approaches, we still believe that there is some potential on investigating methods that leverage differently the contribution of small excerpts of audio or text chunks for the global prediction of emotional state in long narratives.

## 4. Experiments and Results

### 4.1. Experimental set-up

The data set used in this work is the official corpus released for the Elderly Emotion Sub-Challenge of the INTERSPEECH 2020 ComParE Challenge, comprising 261 recordings of 87 German-speaking elderly participants. Train, Development, and Test subsets have the same number of recordings. For each speaker, there are three different narratives, which are labelled regarding their Arousal and Valence as High, Medium, or Low. The recordings of these narratives have been split into smaller chunks, and those are the audios we have access to. Therefore, to find a global label over all the chunks, we had to adopt the decision strategies in Subsection 3.4. We also have transcriptions of the full text, but without correspondence to each of the chunks of the same narrative. In train, the distribution of the labels of Valence is {L:33, M:30, H:24}, whereas for Arousal it is {L:13, M:44, H:30}, which is why it is subject to upsampling before training the classifier. The metric adopted by the Challenge is the Unweighted Average Recall (UAR).

The best results of the official baseline for the development set are UAR = 42.0% for Arousal and UAR = 56.1% for Valence. These were obtained with Bag-of-Audio-Words [23] + SVM and a Linguistic Feature extractor + SVM, respectively [1].

### 4.2. Audio modality results

Table 1 reports Arousal and Valence UAR classification results obtained with each acoustic feature set for the corresponding best SVM complexity configuration. The Table also includes results for the two approaches considered for attributing a single label to a sequence of items: majority voting (MV) and feature averaging (AVG).

Table 1: *Best UAR [%] in development of each individual acoustic feature set for the majority voting (MV) and feature averaging (AVG) approaches.*

| | **Arousal** | | **Valence** | |
|---|---|---|---|---|
| | MV | AVG | MV | AVG |
| ComParE (CP) | 39.1 | 42.7 | 45.7 | 43.4 |
| eGeMAPS (eG) | 42.5 | 44.3 | 36.0 | 40.5 |
| on-line (oXv) Xv | **44.7** | 39.3 | 47.5 | **48.8** |
| normalized (nXv) Xv | 35.8 | 33.3 | 40 | 43.5 |

Table 2 shows the results obtained for the different combinations of hand-crafted and embedding acoustic feature sets.

Table 2: *Best UAR [%] in development of combined hand-crafted and embedding acoustic feature sets for the majority voting (MV) and feature averaging (AVG) approaches.*

| | **Arousal** | | **Valence** | |
|---|---|---|---|---|
| | MV | AVG | MV | AVG |
| CP + oXv | 41.0 | 43.1 | 48.3 | 50.0 |
| eG + oXv | 44.7 | 44.1 | 49.8 | 48.4 |
| CP + nXv | 39.3 | **46.1** | **51.1** | 49.5 |
| eG + nXv | 41.8 | 40.5 | 46.7 | 43.8 |
| CP + sXv | 41.7 | 44.2 | 48.0 | 48.7 |
| eG + sXv | 40.1 | 37.0 | 46.9 | 46.4 |

Amongst each acoustic feature set, Table 1, on-line x-vectors provided the best results. As for the remaining acoustic feature sets, it is hard to find general tendencies on which perform best. As well, no voting approach seems to be better. For the combination of acoustic feature sets Table 2, the best results for Arousal and Valence are obtained for the ComParE set and normalized x-vectors, with feature averaging in one case, and majority voting in the other, although no feature combination nor averaging approach seems to be better than another. It

is noticeable, though, that the combination of different feature vectors steadily improves the performance.

Furthermore, although the number of features in each set may vary considerably, it seems to have no direct impact on the classification.

### 4.3. Text modality results

To keep the size of the input to BERT within its natural limits, we have extracted the embeddings for sentence-like units of each text. The embedding corresponding to the narrative is the average embedding over all the embeddings. The best results of the SVM classification of these embeddings on the development set are UAR = 40.6% for Arousal and UAR = 58.8% for Valence.

Whereas for Arousal the UAR is still under the baseline, for Valence it is already above. This matches the principle that text has a higher contribution to Valence than acoustics.

Linguistic modelling of the baseline was initially obtained from a German BERT model. The best result for Arousal – 40.6% – is the same as ours, but the best result for Valence – 56.1% – is slightly under what we were able to achieve with the multilingual model. Although a direct comparison cannot be made, as the pipelines are different, we can presume that a multilingual model was a good choice, for this case.

### 4.4. Early and late fusion results

To combine acoustic and text systems at the feature level (early fusion), we take the average audio features over all chunks and concatenate them with the corresponding text features for the same narrative. Results are shown in the left-hand side of Table 3. To combine acoustic and text systems after classification (late fusion), we take the classification confidence score of the individual systems (in particular, the ones using the average combination of Tables 1 and 2 and the text embedding system of section 4.3) and, for each narrative, we chose the label either from text or acoustic classification with the highest confidence score. The results are on the right-hand side of Table 3.

Regarding early fusion, differences among the acoustic feature sets for each dimension of emotion are no longer as noticeable as they were in Table 1. We notice that results are leveraged for both dimensions, but more noticeably for Valence. For Arousal, the combination with text embeddings allowed for five combinations surpassing the baseline for the development set, against three for Valence. Regarding late fusion, when compared to the acoustic-only systems in Table 2, we see that although the contribution of text for Arousal is not clear, as there is only one situation where it improves, results tend to be better due to the combination with text modality. For Valence, not only it improves the classification in all cases, but it also seems to equalize them. Early fusion does not lead to this homogeneity, but on the other hand, it allows for performance spikes in some specific configurations, achieving the best results reported in this work.

Overall, the best results for Arousal (48.8%) and Valence (61.0%) are obtained with early fusion of the ComParE feature set and x-vectors – on-line and normalized, respectively. These results correspond to 16.2% and 8.7% relative UAR improvement with respect to the official baseline in the development set. Table 4 shows the confusion matrix of this best system. There, we notice that the system for Arousal classification tends to misclassify the Low and High samples, mostly as Medium. For Valence, the confusion matrix is more balanced. Low samples are the ones in which the system performs best. In both cases, though, there is a bias towards low values, in the sense that misclassifications tend to happen more in attributing labels of lower values than the real one than the opposite.

Table 3: *Best UAR [%] for Early Fusion (left) and Late Fusion (right) of acoustic and text features in development.*

| | **Early Fusion** | | **Late fusion** | |
|---|---|---|---|---|
| | Arousal | Valence | Arousal | Valence |
| (CP) + TE | 46.4 | 52.8 | 40.6 | 48.2 |
| (eG) + TE | 45.1 | 53.3 | 44.0 | 54.1 |
| (CP + oXv) + TE | **48.8** | 56.9 | 43.0 | 53.8 |
| (eG + oXv) + TE | 36.9 | 57.5 | 35.7 | 54.6 |
| (CP + nXv) + TE | 43.9 | **61.0** | **45.4** | 49.1 |
| (eG + nXv) + TE | 40.5 | 43.9 | 40.7 | 56.4 |
| (CP + sXv) + TE | 48.1 | 51.6 | 43.7 | 53.6 |
| (eG + sXv) + TE | 36.9 | 52.1 | 36.5 | **58.0** |

Table 4: *Confusion matrices of the best systems for Arousal and Valence. Entry in row i and column j indicates the number of samples with true label being i class and predicted label being j class.*

| | Arousal | | | Valence | | |
|---|---|---|---|---|---|---|
| | **L** | **M** | **H** | **L** | **M** | **H** |
| **L** | 5 | 12 | 1 | 23 | 9 | 8 |
| **M** | 6 | 33 | 11 | 6 | 13 | 9 |
| **H** | 1 | 8 | 10 | 1 | 3 | 15 |

## 5. Discussion and Conclusions

Emotions can be interpreted as combinations of Arousal and Valence on different levels of granularity. Each of these dimensions may be characterized by different types of features. In this work, we have explored how audio and text embeddings can contribute to the classification of different dimensions of emotions. In particular, we experimented several combinations of traditional paralinguistic-tailored features with model-based features, as well as different early and late fusion schemes for the two modalities. The speech embeddings were x-vectors extracted from a network trained for speakers of a wide range of backgrounds, accents of English and ages, and the linguistic model for BERT was trained on 104 languages. We have seen that these embedding features, which have not even been trained for emotion recognition, can attain – isolated – results comparable or even better to those obtained with traditional features, showing the success of feature embeddings for transfer learning.

To what extent context is important for this task, and particularly fine-grained to this specific population, it is not yet well known. However, it is clear that context/embeddings with different modalities have a role to play in emotion recognition, not restricted to negative emotions of the elderly (usually, the spectrum analysed in the literature).

For future work, it would be interesting to assess how this approach would perform for data of non-elderly and to explore the particularities of this specific population.

# 6. References


[1] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *Proceedings of Interspeech*, Shanghai, China, September 2020, p. 5 pages, to appear.

[2] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[3] J. A. Russel, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, pp. 1161–1178, 1980.

[4] R. Trnka, A. Lačev, K. Balcar, M. Kuška, and P. Tavel, "Modeling semantic emotion space using a 3d hypercube-projection: an innovative analytical approach for the psychology of emotions," *Frontiers in psychology*, vol. 7, p. 522, 2016.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[7] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *arXiv preprint arXiv:2002.05039*, 2020.

[8] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[9] B. Mitra and N. Craswell, "Neural text embeddings for information retrieval," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 813–814.

[10] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] C. Huang, A. Trabelsi, and O. R. Zaïane, "Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert," *arXiv preprint arXiv:1904.00132*, 2019.

[14] S. G. Karadoğan and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," in *2012 3rd International Workshop on Cognitive Information Processing (CIP)*. IEEE, 2012, pp. 1–6.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[16] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[20] J. Rownicka, P. Bell, and S. Renals, "Embeddings for dnn speaker adaptive training," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 479–486.

[21] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[22] H. Xiao, "bert-as-service," https://github.com/hanxiao/bert-as-service, 2018.

[23] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, 2017.