Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), pages 91–97
Marseille, 24 June 2022

Annotation of expressive dimensions on a multimodal French corpus of political interviews

Jules Cauzinille, Marc Evrard, Nikita Kiselov, Albert Rilliard
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)
Campus universitaire – bat. 507 Rue du Belvedère – F-91405 Orsay cedex
{jules.cauzinille, marc.evrard, nikita.kiselov, albert.rilliard}@lisn.upsaclay.fr

Abstract

We present a French corpus of political interviews labeled at the utterance level according to expressive dimensions such as Arousal. This corpus consists of 7.5 hours of high-quality audio-visual recordings with transcription. At the time of this publication, 1 hour of speech was segmented into short utterances, each manually annotated in Arousal. Our segmentation approach differs from similar corpora and allows us to perform an automatic Arousal prediction baseline by building a speech-based classification model. Although this paper focuses on the acoustic expression of Arousal, it paves the way for future work on conflictual and hostile expression recognition as well as multimodal architectures.

Keywords: expressive speech, affective computing, automatic prediction of Arousal, political speech processing

1. Introduction

This paper presents a new corpus of French political interviews annotated on expressive dimensions. We conducted a primary annotation on a subset of a corpus consisting of 7.5 hours of automatically transcribed audio-visual recordings of French broadcast programs. Arousal levels were manually annotated on speech segments. A baseline models for speech-based automatic Arousal prediction is also proposed.

In this paper, we mainly focus on defining annotation guidelines, studying our preliminary labeling, and setting the future work that will be carried out on the complete corpus. We first present the specific characteristics of this political corpus and make a literature review on similar projects. We then describe the methodology of our preliminary annotation and baseline model in Section 2 and showcase the first results in Section 3. In Section 4, we finally discuss these results and their implications for future work.

1.1. Political Scope and Context

The specific nature of political talks, interviews, and speeches makes them an appealing topic for researchers in affective computing and expressive speech processing. Such corpora provide the advantages of studio-quality audio while being expressively diverse and overcoming the limitations of read and acted speech. Professional speakers, such as politicians, produce a particular type of speech that could be described as semi-spontaneous, in the sense that displayed expressive behaviors are part of the politician's communication strategies without being completely scripted and enacted.

This political context also presents a limited variety of expressions. Some projects focus on conflictual and hostile discourse (Kouklia, 2019), arguing that it is the most frequent expressive style in political debates, thus offering a prolific ground for the study of affective expressivity linked to controlled aggression, cold anger, dispute, and hostility.

In this work, we focus on a dimensional model of expressivity, which is inspired by Russel's psychological *circumplex* model of affect (Posner et al., 2005) and Barrett's theory of *constructed emotion* (Barrett, 2016), by Gussenhoven's biological codes (Frequency and Effort codes) (Gussenhoven, 2004), and by the dimensional description of affective meaning proposed by Osgood (Osgood et al., 1975). Although the final goal of this project would be the study of conflicting expressive behaviors in spoken interactions, here we present a preliminary annotation of Arousal, seen as the amount of energy physically displayed in speech and movement. Valence labeling was left for future work as many difficulties arose during the annotation of this dimension in this political context.

1.2. Previous Work

Although our corpus was annotated with a focus on expressive labeling, it resembles datasets belonging to the affective computing domain in general. The majority of such resources are in English, with widely cited affective corpora such as IEMOCAP (Busso et al., 2008) and RAVDESS (Livingstone and Russo, 2018)—both multimodal, acted, and labeled with categorical emotions. Closer to our work would be the large MSP-Podcast dataset (Lotfian and Busso, 2019) based on semi-spontaneous speech from radio shows and podcasts. On top of categorical emotion labels, it presents annotations along the Arousal and Valence dimensions. For French, some of the most used and cited affective corpora are RECOLA (Ringeval et al., 2013) and AlloSat (Macary et al., 2020). RECOLA is a continuously annotated corpus of spontaneous speech with Valence and Arousal labeling for 46 speakers. It gives an insight into the difficulties and limitations of such affective annotations. A major difference between our corpus and RECOLA is its focus on an affective definition of Arousal and Valence compared to our expressive point of view. We rely on a definition of Arousal based on speech production and the concept of vocal effort that may be linked to Gussenhoven's *Effort Code* (Gussenhoven, 2004), or Liénard's definition of vocal effort (Liénard, 2019). Another difference is the context of the spoken interactions, as RECOLA consists of video conference recordings of individuals performing collaborative tasks, while we target broadcast political interviews. RECOLA also proposes a continuous annotation framework while we performed our annotations on discrete segments. We expect that considering entire speech segments allows for a more consistent labeling process, as such annotations are complex and may require several passes over a segment for the annotator to make a decision. More information about the segmentation process can be found in Section 2.

A notable work on French political expressive speech was carried out by Kouklia (Kouklia, 2019) on debates recorded in a French town hall in 2013. The author presented an extensive survey of affective expression in a political context with prosodic analyses carried out on a corpus of annotated speech. It showed the importance of studying affective expressions in political speech for many research areas beyond affective computing and speech processing, while stressing the number of perspectives opened up by the annotation of such data.

2. Methods

2.1. Annotation Framework

The complete corpus consists of 7.5 hours of political interviews from two French television channels: *BFM TV* and *France 2*. It includes 30 different speakers—and 4 interviewers—for an average of 20 minutes per interview for *BFM TV* and 8 minutes for *France 2*. These sequences contain video, studio-quality audio, an automatic transcription aligned at the word-level, and manually annotated speech turns and overlaps. We present an annotation on a subset of the whole corpus: 5 minutes extracts from 12 different interviews for a total of 1 hour. This work focuses on the Arousal dimension although an attempt was made at annotating Valence.

The annotation was carried out with the *PRAAT* and *ELAN* software programs. *PRAAT* allowed speeding up the segmentation task using only the audio and its transcription, while *ELAN* allowed for the use of visual information to perform a genuinely multimodal expressive annotation.

2.1.1. Segmentation

Most available datasets labeled on similar dimensions, such as RECOLA (Ringeval et al., 2013), were annotated in a time-continuous way, resulting in evolving Arousal and Valence trends spanning entire sound files. In other datasets, such as in Kouklia (2019), single labels will be assigned to speech utterances that can often be quite long, possibly yielding multiple emotional and expressive states within a single annotated segment.

In this regard, we first tried to settle on a syntactic segmentation. Because, contrary to written text, speech transcriptions cannot generally be segmented into sentence-like units with clear punctuation boundaries (Ostendorf et al., 2008), we identified short syntactic units consisting of several words that could be isolated as meaningful syntagms, usually separated by discourse markers or pauses.

Unfortunately, this approach tends to produce segments of various sizes, which is unpractical both for the annotation process and for the implementation of computational models.

With this in mind, we adapted our segmentation approach by following three criteria: a boundary is added either when the annotator perceives a significant expressive variation, when a semantically meaningful unit can be found, or when the segment exceeds a fixed threshold. We found that a duration of about 3 seconds was optimum for the annotation task (resulting in a median of 2.14 seconds per segment).

2.1.2. Expressive Dimensions

As previously said, we adopted Arousal and Valence dimensions as the main variables of vocal expression. Our definition is inspired by studies emphasizing the role of Arousal in the vocal expression of emotions (Goudbeek and Scherer, 2010), and the importance of vocal effort as a characteristic of the voice (Liénard, 2019; Titze and Sundberg, 1992). We adapted this, as well as previously cited psychological models, to serve the concept of Arousal as a vocal feature. In this respect, it should be seen as the degree of vocal effort and energy displayed by a speaker along a given utterance. It is cued at different levels within speech: on syllable-sized segments, short utterances (which correspond to our approach), or longer units of discourse. We broadened Rilliard et al. (2018)'s description of Arousal on syllable-sized units to our longer segments. In this perspective, low Arousal is characterized by a slow speech rate, low intensity, and steady fundamental frequency due to slower vibration of the vocal folds. Low Arousal may also lead to specific voice qualities such as creaky and breathy voices. The high energy Arousal, on the contrary, is cued by greater variability in speech rate and f_0, large f_0 span, high intensity levels, and an increased overall vocal effort. Of course, most of these vocal events may not be directly linked to an Arousal level, which is instead signaled by a combination of several features. For instance, although they can be seen as typical cues of Arousal, speech rate and pause frequency or duration may fail to show a strict correlation with it, especially in the context of political speech (Madureira and de Camargo, 2019; Kouklia, 2019).

After experimenting with different options for the description of levels of Arousal, we settled on a 7-level Likert scale (Joshi et al., 2015) from -3 to 3, although both minimal values (-3 and -2) were never encountered in our annotation process. With neutrally aroused speech—0 in such a scale—usually being the minimally aroused expressive style in a political context, levels corresponding to truly sleepy, extremely depressed, and underactivated speech are, as expected, barely used in our annotation. Political expression is indeed typically characterized by positive Arousal, as shown by Vázquez et al. (2019). We still kept a 7-level scale, including negative Arousal, for better generalization of our annotation method.

A remaining question regarding this scale is to decide on the neutral Arousal level for each speaker, as some of them tend to be more or less active when they speak with seemingly neutral expressivity. We answered this problem by asking annotators to base their labels on a common neutral Arousal value of 0 (which would correspond to a "typical" speaker) and annotate speakers showing higher than normal neutral speech styles with corresponding higher levels. This common scale applies to every speaker and allows building computational models without the additional difficulty of predicting speaker-dependent labels.

As we previously mentioned, we attempted to annotate vocal Valence. Unfortunately, compared to the expression of Arousal, Valence is defined by a very complex set of acoustic features and is considered, as shown by Belyk and Brown (2014), to be highly dependent on emotional contexts, which are not the main focus of our annotation. It is also a more variable dimension, as for its acoustic characteristics (Goudbeek and Scherer, 2010), typically dependent on Arousal. In addition to that, the political context of our corpus makes Valence labeling particularly difficult to grasp (Vázquez et al., 2019). We still conducted a preliminary Valence annotation to get an idea of the statistical distribution of values over part of the corpus.

2.2. Baseline Models

Based on our annotations, we built two baseline models for automatic vocal Arousal classification. The first consists in using the self-supervised learning framework wav2vec 2.0 (Baevski et al., 2020) for feature extraction (encoding each segment of speech into a fixed size matrix embedding). A subsequent Gated Recurrent Unit (GRU) architecture was then trained on top of these representations to predict the arousal values.

The second consists in a Convolutional Neural Network (CNN) trained on top of Mel-Frequency Cepstral Coefficients (MFCC) representations of each segment. This model serves as a comparison with the pre-trained approach and allows us to test if the corpus contains enough information to train a simple neural network on spectral features.

2.2.1. Wav2vec-Based Models

The model is based on pre-trained wav2vec 2.0 feature extraction from the multilingual *facebook/wav2vec2-large-xlsr-53* (Conneau et al., 2020). We also tested a french version, *LeBenchmark/wav2vec2-FR-1K-large* (Evain et al., 2021), which showed significantly lower performances. The wav2vec 2.0 representations were extracted with the *Huggingface* module and all models are built with *Pytorch*. All feature matrices were trimmed and zero-padded to obtain a segment size of 3 seconds, corresponding to embeddings of size (150×1024), with 50 wav2vec features of size 1024 for each second. The best performing model was a GRU trained on these sets of features. It is built with one layer, a hidden size of 128, sigmoid activation, and 10% dropout.

2.2.2. MFCC-Based Model

A set of 13 MFCCs was extracted for each segment with the *Librosa* package. The segments were padded or trimmed to be 120 frames long (40 per second for 3 seconds segments). They were subsequently processed by a CNN with three 2D convolutional layers and three linear layers.

The best-performing CNN's architecture and hyperparameters were inspired by models found in the literature, such as Zhao et al. (2019), and through empirical testing.

Each layer consists in a convolution kernel of size (3×3) with a ReLU activation function and, respectively, 64 and 128 filters. Two (2×2) max-pooling layers are added after each convolution. Three fully connected linear layers are then applied with 30% dropout.

3. Results

3.1. Inter-Annotator Agreement

Inter-annotator agreement may be assessed in different ways for expressive dimensions labeling. Considering a strict correspondence of each segment with a defined *class* is not coherent with the *degrees* of an Arousal scale, and the distance between these degrees should be taken into account when measuring the annotator's agreement.

Although more than two annotators would be required to build a robust annotation, our preliminary labeling shows promising results when agreement is tested through the quadratic-weighted Kappa (Artstein and Poesio, 2008) metrics: $\kappa_w = 0.546$. This value was obtained by concatenating the annotations of all interviews. The resulting Kappa score may be considered as a *moderate* agreement.

3.2. Annotation and Segmentation Distribution

A first step in investigating our preliminary corpus is to study the distribution of annotated values in order to describe Arousal dynamics. The reported values are the average of those given by both annotators. Over the whole dataset, Arousal tends to be positive, with an average of 0.7 (on the $-3/+3$ scale) for the 12 interviews. As expected in the context of political speech (Vázquez et al., 2019; Kouklia, 2019), Arousal distribution is dominated by positive values: 51.2% of segments showed an Arousal between 0 and 1, 41.4% between 1 and 2, and only 5.3% from 2 onward. Negative values are relatively rare as only 2.1% of the data was labeled below 0, and no segments were given a value below -1. The global distribution of the Arousal levels can be seen in Table 1 and Fig. 1. Violin plots for each interview are shown in Fig. 2.

Arousal interval	Count	Percentage
$[-1.25, -0.75)$	3	0.18%
$[-0.75, -0.25)$	33	1.94%
$[-0.25, 0.25)$	465	27.37%
$[0.25, 0.75)$	408	24.01%
$[0.75, 1.25)$	503	29.61%
$[1.25, 1.75)$	202	11.89%
$[1.75, 2.25)$	70	4.12%
$[2.25, 2.75]$	15	0.88%

Table 1: Distribution of Arousal levels for all segments in the dataset. Arousal values are the average of those given by both annotators.

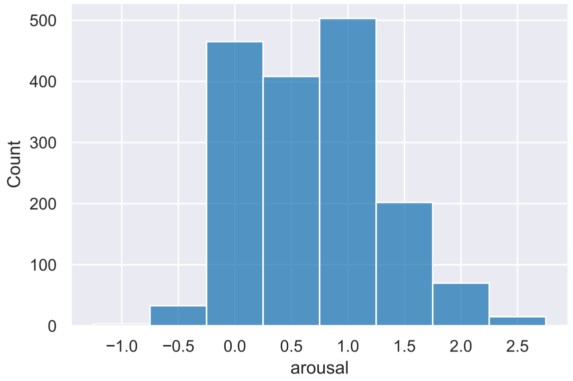

Figure 1: Distribution of Arousal levels for all segments in the dataset. Arousal values are the average of those given by both annotators.

Regarding the segmentation process, the median segment length is 1.81 seconds (see Fig. 3), and 97% of them are under the 3 seconds threshold that we defined as the upper bound.

3.3. Baseline Results

The performance scores obtained by our two models can be seen in Table 2. They were computed on the entire annotated dataset through a 12-folds cross-validation procedure (each fold containing 11 interviews for the training set and 1 for the test set). This allowed us to obtain a representative performance and to limit the bias implied by the small size of the corpus, as testing the models on different interviews may yield different results.

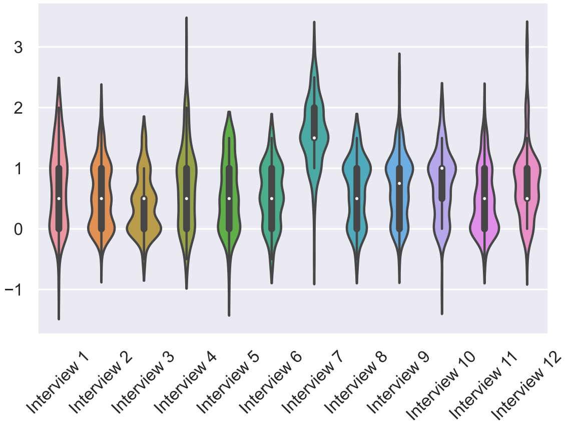

Figure 2: Violin plots of Arousal levels for segments in the 12 interviews. Arousal values are the average of those given by both annotators.

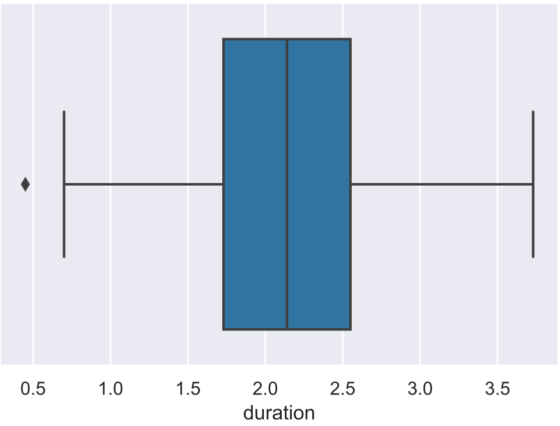

Figure 3: Distribution of segments lengths in seconds.

In addition to the mean squared error (MSE) that was used as a loss function for the training process, which yielded better results than L1 loss, we computed traditional regression metrics such as root mean square error (RMSE) and mean absolute error (MAE).

Model	RMSE	MSE	MAE
MFCCs+CNN	0.555 (0.064)	0.322 (0.081)	0.464 (0.053)
Wav2vec+GRU	0.577 (0.062)	0.336 (0.073)	0.461 (0.051)

Table 2: Mean results for automatic Arousal prediction on the 12-folds cross-validation, with the standard deviation in brackets.

4. Discussion

The general trend of the annotation is quite similar between both annotators, with the majority of differences rarely exceeding one degree on the scale, as shown by shaded area heights in Fig. 6 and 7. The obvious solution to merge both annotations was to compute their average. In fact, if a segment is labeled higher by one annotator and lower by another, then the corresponding level of Arousal is most certainly ambiguous and should be averaged to an intermediary level. This observation also raises the important question of soft-labeling that may be applied to this type of annotation. The idea would be to take into account a certainty degree on each annotation value in order to address the problem of variability in subjective annotations. We leave these considerations for future work.

With regard to the distribution of Arousal and Valence labels, we observed good correspondence with our first intuition, as well as previous work carried out by (Vázquez et al., 2019; Kouklia, 2019). Expressive political speech shows a tendency to express conflict and hostility, denoted by these generally positive values of Arousal and negative values of Valence. In addition to these metrics, it is also interesting to compare speakers' behaviors, keeping in mind that a 5 minutes annotation is not perfectly representative of a one's expressive style. As shown in Fig. 4 and 5., interviews considered to be more conflictual by the annotators exhibit significantly different distributions when compared to the more neutral ones.

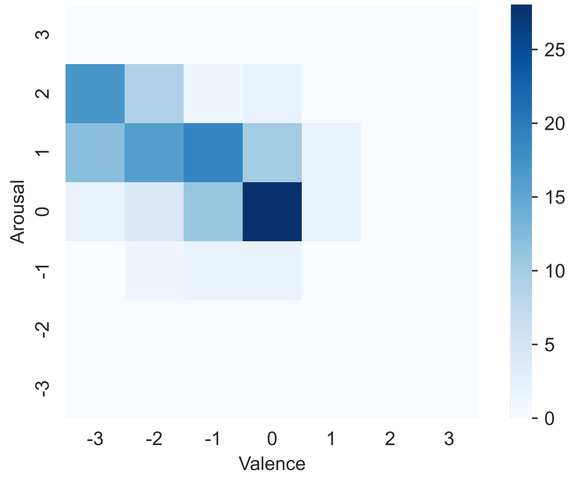

Figure 4: Heatmap of the Arousal-Valence distribution for a "conflictual" interview (Interview 7).

Finally, one can observe the variation of Arousal levels throughout a given extract, as in Fig. 6 and 7.

Regarding the automatic prediction of Arousal levels from speech, the obtained results showed that convergence is possible but that more work needs to be carried out on a variety of architectures to handle the task thoroughly. Both the Wav2vec transfer learning approach and the MFCC-based CNN showed promising performances given the scarcity of data on which they were trained. Fig. 8 shows the similar tendencies of predictions and labels in one of the interviews.

Although very different in their architecture and feature extraction process, both models exhibit very similar performances (see Table 2), with differences within the margin of error. This tends to confirm that we may have reached an upper limit on the dataset in its current state.

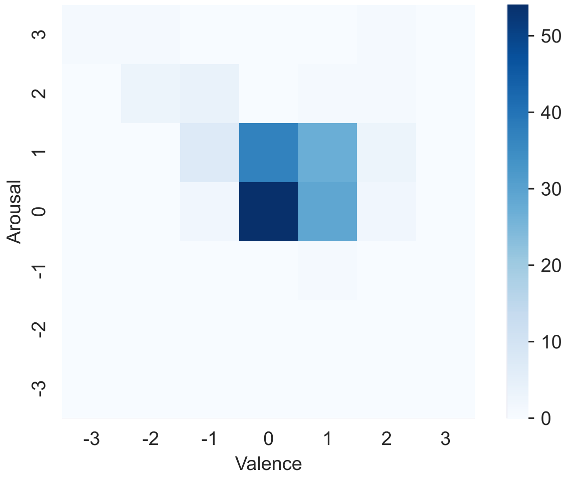

Figure 5: Heatmap of the Arousal-Valence distribution for a "neutral" interview (Interview 4).

5. Conclusion

In this work, we discussed the challenge of creating a stable and coherent annotation framework for political interviews. We have seen that dimensional labels, such as the amount of expressed Arousal, are well suited for labeling political speech expressivity and can lead to future annotation and implementation of more specific speech-based conflict recognition models. We also argued that a discrete approach with segmented extracts of speech, each labeled on Arousal, is an effective way to ease the annotation process and allow for efficient training of computational models.

We presented a preliminary corpus annotated on the Arousal dimension for 12 speakers and conducted several statistical experiments to describe the annotation distribution. Finally, we proposed two regression models, exploiting wav2vec 2.0 feature extraction and MFCC-based architectures as a first baseline for automatic Arousal prediction.

After validating our annotation process, we plan on performing it on the entirety of the 7.5-hour corpus with a higher number of participants. The complete annotation would also benefit from an extended study of the Valence dimension and a more precise definition of its annotation. It will also include a new label, dependent on the ones we already discussed, exploring conflict

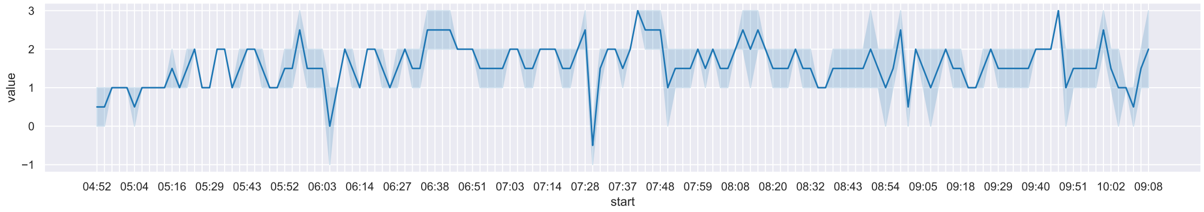

Figure 6: Averaged Arousal from both annotators for a typical "conflictual" interview (Interview 7). The shaded area shows the differences between values chosen by the annotators.

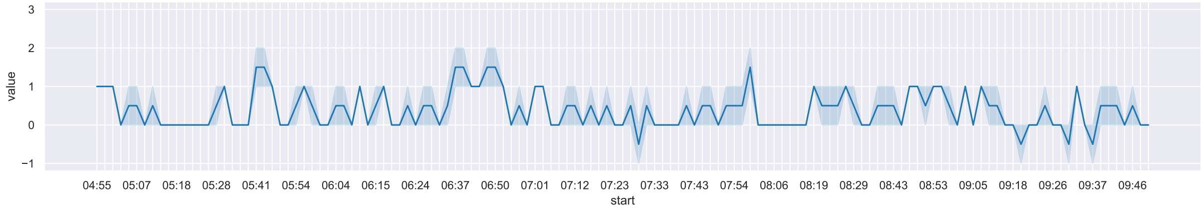

Figure 7: Averaged Arousal from both annotators for a typical "neutral" interview (Interview 3). The shaded area shows the differences between values chosen by the annotators.

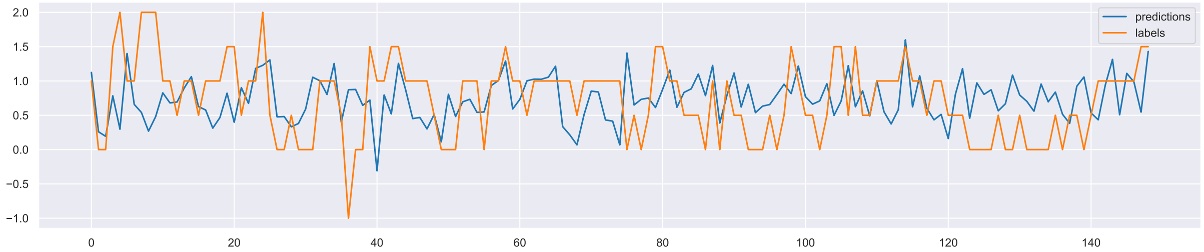

Figure 8: Visual representation of predictions and labels for the MFCC-based CNN (Interview 10).

versus complicity expression or the degree of approval and hostility between speakers.

6. Acknowledgments

This research was supported by the DATAIA/MSH Paris-Saclay "Excellence" grant "OTELO", and was part of a cooperation of the CESSP (Paris Panthéon-Sorbonne University, CNRS), the Institut national de l'audiovisuel (INA), and the LISN (Université Paris-Saclay, CNRS) through the OOPAIP project under the DIM STCN (Ile de France). The corpus was initially built with the OKAPI platform developed by INA, thanks to the work of Steffen Lalande and Abdelkrim Beloued.

7. Bibliographical References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Barrett, L. F. (2016). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 10.

Belyk, M. and Brown, S. (2014). The acoustic correlates of valence depend on emotion family. *Journal of Voice*, 28(4):523.e9–523.e18.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Evain, S., Nguyen, H., Le, H., Boito, M. Z., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., et al. (2021). Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech. In *INTERSPEECH*.

Goudbeek, M. and Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128:1322–36, 09.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge University Press, 07.

Joshi, A., Kale, S., Chandel, S., and Pal, D. K. (2015). Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Kouklia, C. (2019). *Dominance, hostilité et expressivié vocale dans le débat politique: étude perceptive et acoustique du conseil municipal de Montreuil (93100)*. Ph.D. thesis, Université Sorbonne Paris Cité.

Liénard, J.-S. (2019). Quantifying vocal effort from the shape of the one-third octave long-term-average spectrum of speech. *The Journal of the Acoustical Society of America*, 146(4):EL369–EL375, October.

Livingstone, S. and Russo, F. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13:e0196391, 05.

Lotfian, R. and Busso, C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October-December.

Macary, M., Tahon, M., Estève, Y., and Rousseau, A. (2020). AlloSat: A New Call Center French Corpus for Satisfaction and Frustration Analysis. In *Language Resources and Evaluation Conference, LREC 2020*, Marseille, France, May.

Madureira, S. and de Camargo, Z. A. (2019). Exploring sound symbolism in the investigation of speech expressivity. *International Speech Communication Association*, page 105.

Osgood, C. E., May, W. S., and Miron, M. S. (1975). *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, Baltimore, MD, June.

Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tur, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., and Woofers, C. (2008). Speech segmentation and spoken document processing. *Signal Processing Magazine, IEEE*, 25:59 – 69, 06.

Posner, J., Russell, J., and Peterson, B. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17:715–34, 02.

Rilliard, A., d'Alessandro, C., and Evrard, M. (2018). Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis. *The Journal of the Acoustical Society of America*, 143(1):109–122.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, pages 1–8, 04.

Titze, I. R. and Sundberg, J. (1992). Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America*, 91(5):2936–2946, May.

Vázquez, M. d., Justo, R., Zorrilla, A. L., and Torres, M. I. (2019). Can spontaneous emotions be detected from speech on tv political debates? In *10th IEEE International Conference on Cognitive Infocommunications*, page 289.

Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323.