Emotion recognition from phoneme-duration information

Ajinkya Kulkarni[1] , Ioannis K. Douros[1,2], Vincent Colotte[1], Denis Jouvet[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France,
[2]Université de Lorraine, INSERM U1254, IADI, F-54000 Nancy, France

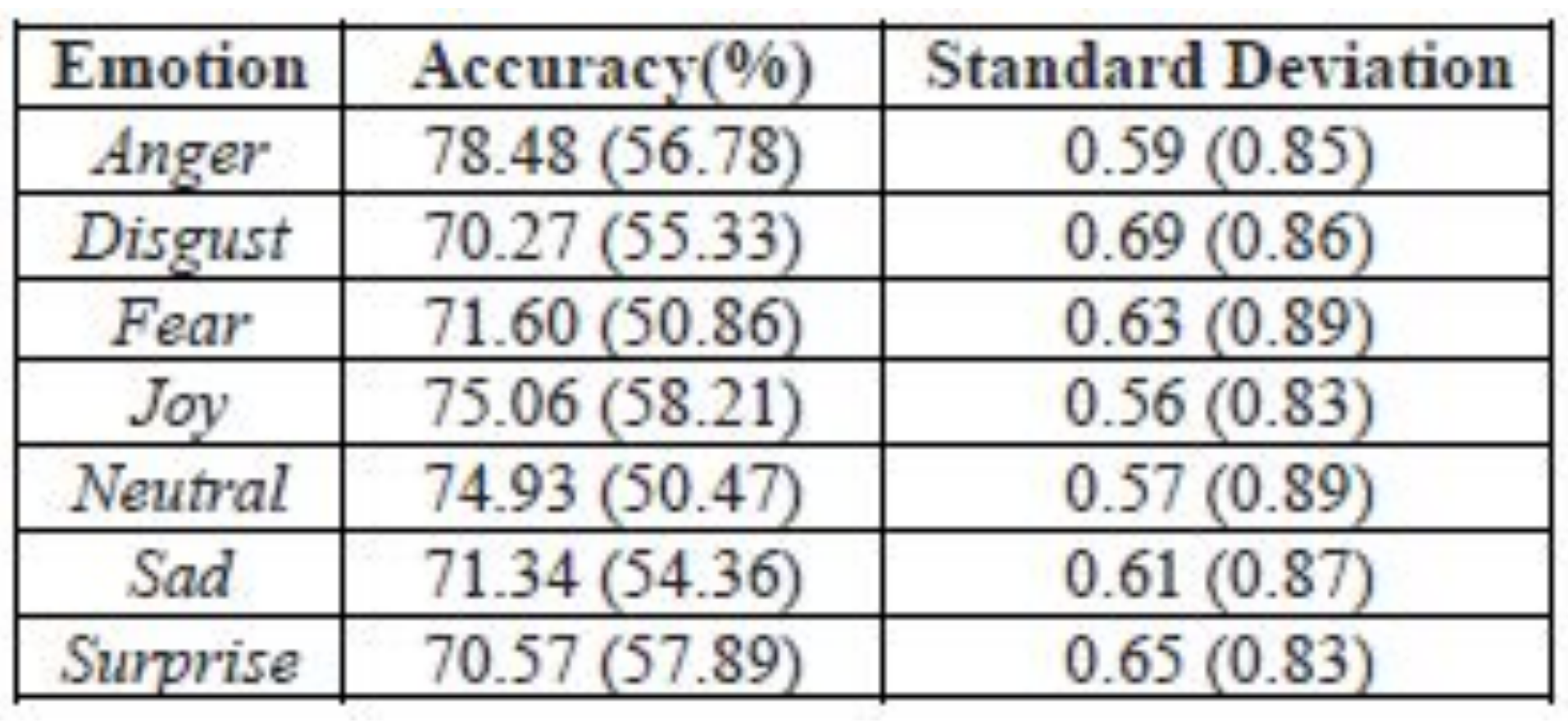
ajinkya.kulkarni@loria.fr, ioannis.douros@loria.fr, vincent.colotte@loria.fr, denis.jouvet@inria.fr

Objectives

The main purpose of the presented algorithm is to study the correlation of phoneme durations with emotions and illustrate the importance of phoneme duration on emotional speech production. Duration of each phoneme is extracted for several emotions. Phoneme information and its duration are used to train a Variational AutoEncoder (VAE) in order to create the latent space z which represents emotion information. The loss functions that were used for that purpose are reconstruction loss, Kullback-Leibler (KL) divergence and multiclass N pair loss. Test samples are classified using nearest neighbour between their representation and the training clusters of the latent space. To evaluate the models two metrics were used: emotion recognition accuracy and the consistency of the clusters of the latent space.

Results

On the Tab. 1 we can see the accuracy for the baseline system using reconstruction and KL divergence as well as the proposed system where we added multiclass N pair loss. The proposed system has an average accuracy of 73.18% compared to 54.84% of the baseline system. We can see that the latent space of the proposed model has more compact clusters (average standard deviation of 0.61 compared to 0.86 of the baseline model) and are clearly separated from each other without overlapping.



Emotion	Accuracy(%)	Standard Deviation
Anger	78.48 (56.78)	0.59 (0.85)
Disgust	70.27 (55.33)	0.69 (0.86)
Fear	71.60 (50.86)	0.63 (0.89)
Joy	75.06 (58.21)	0.56 (0.83)
Neutral	74.93 (50.47)	0.57 (0.89)
Sad	71.34 (54.36)	0.61 (0.87)
Surprise	70.57 (57.89)	0.65 (0.83)

Tab. 1: Result table using VAE with multiclass N pair loss. In parenthesis are the results with the simple VAE approach

Discussion and conclusion

The t-SNE plot of the VAE N-pair model shows well-clustered emotion in latent space. The orange cluster in the t-SNE plot represents neutral speech. From the figure, addition of multiclass N-pair loss clearly indicates improvement in clustering in latent space, which results in improved performance in emotion recognition illustrated by numerical results. Further research could include extending the results to multispeaker experiments or using other metric losses and architectures for building the model.

Materials and Methods

For this task we used Caroline expressive speech corpus recorded in French language with a female voice. Caroline's expressive speech corpus consists of several emotions, namely joy, surprise, fear, anger, sadness, and disgust (approx. 1hr for each emotion and 3hrs for neutral). For each emotion, there are approximately 500 utterances for a total of 1hr duration. All the speech signals were used at a sampling rate of 16 kHz. Each speech corpus is divided into train, validation, and test sets in the ratio of 80%, 10%, 10% respectively. We used Soja tool [2] as a front end for context label generation. The contextual label file for the French language is designed with 1356 questions which consider the linguistic, phonetic and prosodic details about phonemes such as phonetic category, positional information, stress and accent information, guessed part of speech (GPOS) corresponding to the part of speech annotation of words in the text. 76 questions concern the left-left phoneme (i.e. two phonemes preceding the current one), 76 questions concern the left phoneme (i.e. phoneme preceding the current one), and so on. Contextual labels and speech waveforms are forced aligned using HTK toolkit. For the next step, we used VAE trained with phoneme and corresponding duration information. For the VAE architecture, we implemented a BLSTM based encoder network. The input of the encoder is a sequence of context label features, x, along duration information. The activation of hidden states of BLSTM layer is given to feedforward layers to estimate both mean vector and variance vector, which are used to describe the encoder's latent variable, z. Similarly, the decoder network consists of BLSTM layers. The usage of BLSTM based recurrency allows the model to extract long term context from phoneme and duration information. The input of the decoder network is the latent variable z. The decoder generates the sequence of predicted duration x̂. In inference phase, we provide duration information and context labels as input to encoder of VAE model to generate the latent representation of given speech utterance. Afterwards, we compute the distance between precomputed means of each emotion and generated latent representation. We classify the emotion for given speech utterance considering the minimum distance between precomputed means for emotion. Thus, for better emotion recognition desired latent space should have well separated cluster's corresponding to the various emotions. Therefore, we proposed to use multiclass N-pair loss in variational inference as deep variational metric learning. Multi-class N-pair loss has shown superior performance compared to triplet loss or contrastive loss by considering one positive sample and N-1 negative samples for N classes [3]

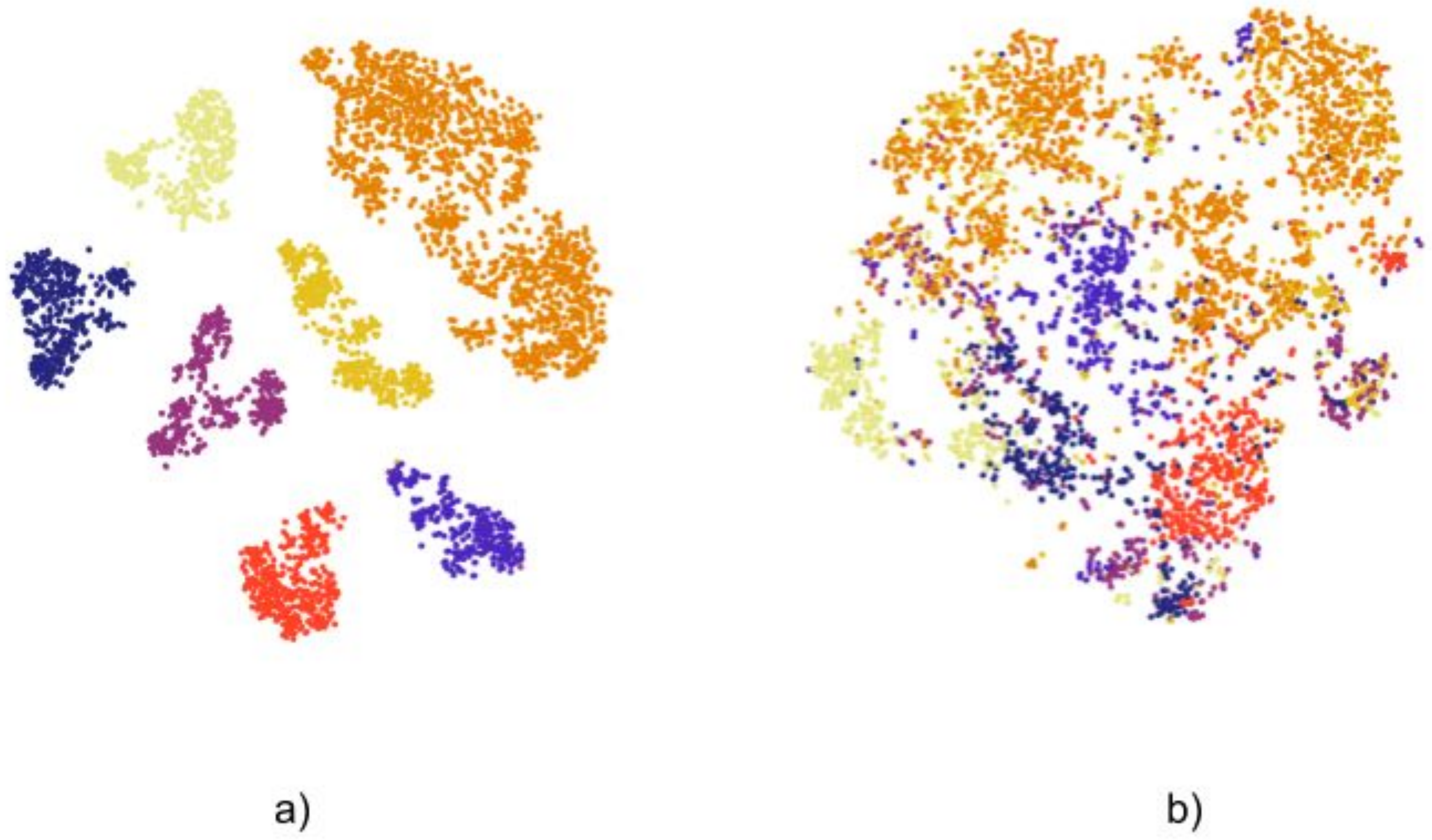


Fig. 1: Latent space z t-SNE representation with N pair loss (a) and without (b).

References

[1] El Ayadi et al. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition(2011), 44(3), 572-587.
[2] https://raweb.inria.fr/rapportsactivite/RA2017/multispeech/uid38.html
[3] Kulkarni, Ajinkya, et al. "Transfer learning of the expressivity using FLOW metric learning in multispeaker text-to-speech synthesis." INTERSPEECH 2020