Analyzing the impact of including listener perception annotations in RNN-based emotional speech synthesis

Jaime Lorenzo-Trueba[1,a] Gustav Eje Henter[1] Shinji Takaki[1] Junichi Yamagishi[1,2]

Abstract: This paper investigates simultaneous modeling of multiple emotions in DNN-based expressive speech synthesis, and how to represent the emotional labels, such as emotional class and strength, for this task. Our goal is to answer two questions: First, what is the best way to annotate speech data with multiple emotions? Second, how should the emotional information be represented as labels for supervised DNN training? We evaluate on a large-scale corpus of emotional speech from a professional actress, additionally annotated with perceived emotional labels from crowdsourced listeners. By comparing DNN-based speech synthesizers that utilize different emotional representations, we assess the impact of these representations and design decisions on human emotion recognition rates.

Keywords: Emotional speech synthesis, recurrent neural networks, speech perception

1. Introduction

In this paper we investigate schemes for the simultaneous modeling of multiple emotions and how to represent the emotional labels such as emotional class. Our goal is to answer the question of what is the best way to annotate speech data with multiple emotions. To answer this question, we compare an emotional one-hot vector that represents a speaker's intended emotional categories with another emotional vector that represents listener perceptions of the emotional contents as additional auxiliary inputs to DNN-based acoustic models.

Second, we investigate how emotional information should be represented as labels for supervised DNN training, e.g., should emotional class and emotional strength be factorized into separate inputs or not? Therefore we compare DNN systems where the perceived emotional information is jointly represented with another system where the emotional information is factorized.

All the comparisons were done by using a large-scale corpus of emotional speech from a professional actress, additionally annotated with perceived emotional labels from crowdsourced listeners.

2. Emotional representations for DNN-based speech synthesis systems

2.1 Discrete representations

2.1.1 Representation based on talker categories

Since speech synthesis normally uses acted emotions, the easiest way to acquire an emotional representation in an acoustic model is to use emotional categories that the speaker intended to express or was instructed to express during voice recordings (which we call ”talker categories”) and represent it on the basis of the standard one-hot vector. The vector may be used as an additional auxiliary input to the DNN-based acoustic models. In our study, we use a RNN with long-short-term memory units (LSTM) [1]. This means that only the i-th element will be set to 1 for the talker's emotion i, and the remaining elements will be 0.

2.1.2 Representation based on listener dominant categories

It is a fact that the way we speak is not constant [2]. Even an acted emotion may be perceived by some of listeners as a different emotion from the one the talker intended to express. Therefore, a natural way of obtaining accurate emotional categorical representations would be to have multiple listeners annotate the emotional categories that they perceive when they listen to emotional speech. We call the categories as *listener emotional categories*.

2.2 Continuous representations

2.2.1 Emotional confusion matrix based on talker categories

The above one-hot vector is a discrete representations of an emotion, although it is claimed that the emotional space is a continuum rather than a discrete space [3]. The consideration of the talker and listener categories results in an emotional confusion matrix, which can be the basis of continuous emotional representations that reflect both categories jointly, which we refer to as ”perception vectors” in this paper. An example of such a confusion matrix can be seen in Table 1, where rows show the talker categories and columns show the listener ones. Here, e_{ij} shows how much talker's i-th emotion may be perceived as the j-th emotion by listeners as a probability.

2.2.2 Representation based on a row of the matrix

If a row of the confusion matrix is used as an emotional category representation, we obtain a continuous vector that represents the *taker categories weighted by listeners perception*, which is a natural extension of the one-hot vector based on talker categories and is represented as $\widehat{e_i} = (e_{ij}, e_{i2}, ..., e_{iC}, e_{iO})$.

[1] National Institute of Informatics, Tokyo, Japan
[2] The University of Edinburgh, Edinburgh, UK
[a] jaime@nii.ac.jp

Table 1 Confusion matrix of talker and listener emotional categories

	Listener categories					
Talker's categories	**1**	···	**j**	···	**C**	**O**
1	e_{11}		e_{1j}		e_{1C}	e_{1O}
i	e_{i1}		e_{ij}		e_{iC}	e_{iO}
C	e_{C1}		e_{Cj}		e_{CC}	e_{CO}

2.2.3 Representation based on a column of the matrix

It is also possible to use a column of the confusion matrix as another continuous emotional category representation. Contrary to the row case, the j-th column represents how much each talker's emotion may be perceived as the j-th emotion by listeners. Therefore, we may re-label an emotional category of each sentence to a new class perceived by dominant listeners, and a vector $\overline{\boldsymbol{e}}_j = (e_{j1}, e_{j2}, ..., e_{jC})$ may be used as a representation of the listener's j-th emotional category.

2.2.4 Emotional confusion matrix based on listeners categories

If we have multiple listeners per utterance, we can also re-annotate the entire database according to the listeners' annotations and generate a new confusion matrix based on listeners categories, which may be an alternative way of representing emotional categories. In this case, both rows and columns of the matrix then represent categories in the listeners' domain including "other" and *shows variations among the listeners' responses*.

2.3 Representations for perceived emotional strength

Some utterances may sound more expressive than others. We cannot assume that all the utterances that are labeled as the same emotional category will always have the same emotional strength. We may annotate the perceived emotional strength per sentence by computing the average across multiple listeners.

3. Experiment

The main objective of the experiment was to compare the modeling accuracy of a different number of emotional modeling strategies. The perceptual evaluation measured emotional strength, and emotion identification rates, although for this study we only considered identification rates.

3.1 Evaluation

We considered the following 12 modeling strategies:

(1) Talker labeling, one-hot vector (w. and w/o. ES)
(2) Talker labeling and categories (w. and w/o. ES)
(3) Talker labeling, listeners categories (w. and w/o. ES)
(4) Listener labeling, one-hot vector (w. and w/o. ES)
(5) Listener labeling, talkers categories (w. and w/o. ES)
(6) Listener labeling and categories (w. and w/o. ES)

3.2 Results

A total of 54 native Japanese speakers took part on the evaluation, for a total of 4200 evaluated utterances.

3.2.1 Modeling accuracy

We measured the modeling accuracy of each representation by obtaining the Frobenius distance between the confusion matrices of the considered emotional representation and the confusion matrix of natural speech (Table 2). In this metric, the shorter the distance the closer we are to representing natural speech.

Table 2 Frobenius distances of the confusion matrices of the evaluated systems to the confusion matrix for natural speech and to the identity matrix. Here, **Vs. Nat** means the distance to natural speech and **Vs. ID** means the distance to the identity matrix.

Inputs	Labeling	Categories	Vs. Nat	Vs. ID
Confusion	**Talker**	**One-hot**	0.89	1.53
		Talker	0.70	1.49
		Listener	0.63	1.41
	Listener	**One-hot**	0.95	1.68
		Talker	0.74	1.42
		Listener	0.75	1.50
Confusion +ES	**Talker**	**One-hot**	0.95	1.59
		Talker	0.75	1.40
		Listener	0.61	1.31
	Listener	**One-hot**	0.81	1.48
		Talker	0.75	1.41
		Listener	0.82	1.54

From the results, we first see how the one-hot vector categories performed significantly worse for both distances, proving that it is significantly helpful for our emotional system to include the perceptual information of the database. Second, we can see how listener labeling does not appear to help in achieving better modeling accuracy. Emotional strength information improved accuracy only when used together with the listener categories. Finally, the best emotional representation in terms of Frobenius distance overall was based on the talker labeling, listener categories, and emotional strength, for a distance of 0.61.

We can also see that the database labeling process based on listener classes did not improve the performance. This may be partially explained by the limited number of listeners used for individual sentences and by the unbalanced distribution of each emotional category after the re-labeling.

4. Conclusions

The evaluation showed how it is best to use emotional labels based on talker intention instead of on listener perception, at least if the re-labeling process is based on a limited number of annotations or if it skews the balance of the training data. Even so, training on listener categories provided better results than talker categories, with one-hot vectors showing the worst modeling accuracy performance. Finally, the evaluation also showed how emotional strength can increase the modeling accuracy for some emotional representations, and more particularly, for the optimum configuration of talker labels with listener categories.

As future work, we want to try controlling the produced expressiveness to see if we are capable of manipulating the perceptual vectors to both enhance and de-enhance the synthesized emotions.

References

[1] Fernandez, R., Rendel, A., Ramabhadran, B. and Hoory, R.: Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional, Deep Recurrent Neural Networks, *Proceedings of Interspeech*, pp. 2268–2272 (2014).

[2] Athanasopoulou, A. and Vogel, I.: Acquisition of prosody: The role of variability, *Speech Prosody 2016*, pp. 716–720 (2016).

[3] Bänziger, T., Patel, S. and Scherer, K. R.: The role of perceived voice and speech characteristics in vocal emotion communication, *Journal of nonverbal behavior*, Vol. 38, No. 1, pp. 31–52 (2014).