# Investigating different representations for modeling multiple emotions in DNN-based speech synthesis

*Jaime Lorenzo-Trueba*[1], *Gustav Eje Henter*[1], *Shinji Takaki*[1], *Junichi Yamagishi*[1,2], *Yosuke Morino*[3], *Yuta Ochiai*[3]

[1]National Institute of Informatics, Tokyo, Japan
[2]The University of Edinburgh, Edinburgh, UK
[3]Toyota Motor Corporation, Shizuoka, Japan

jaime@nii.ac.jp, gustav@nii.ac.jp, takaki@nii.ac.jp, jyamagis@nii.ac.jp, yosuke_morino@mail.toyota.co.jp, yuta_ochiai_aa@mail.toyota.co.jp

## 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 – should we use the labels that the speaker intended to express, or labels based on listener perception of the resulting speech signals? Second, how should the emotional information be represented as labels for supervised DNN training, e.g., should emotional class and emotional strength be factorized into separate inputs or not? 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 and perceived emotional strength.

**Index Terms**:Emotional speech synthesis, deep neural network, recurrent neural networks

## 1. Introduction

Speech synthesis is a technique that is aimed at generating natural-sounding intelligible speech from arbitrary tests. This has been studied for a long time.

Neural network (NN)-based methods are currently being actively investigated because they are able to significantly improve the quality and naturalness of synthetic speech compared with traditional approaches [1]. First, feed forward (FF) deep neural network (DNN) systems have been proposed as a replacement for decision tree approaches in HMM-based speech synthesis, which have demonstrated better proficiency at handling large amounts of speech data than traditional approaches [2]. Long short-term memory (LSTM)-based recurrent neural networks (RNNs) have long been adopted, and it has been reported that they provide even better naturalness and prosody of synthetic speech due to their capability to modeling the long-term dependencies of speech [3].

Lately, we have seen a sharp rise in the number of proposals of techniques based on convolutional neural networks (CNNs) [4], waveform modeling [5], and generative adversarial networks [6]. These approaches aim to solve one of the oldest problems in SPSS: vocoding. By applying waveform-level modeling and bypassing parametrization, there is no need for waveform generation systems, resulting not only in much more natural and higher quality synthetic speech but also much more versatile sound generation systems as generating music or noises becomes possible [5].

We have also seen a number of advances lately that are showing control capabilities for voice aspects such as those by Li and Zen [7], who built multi-language and multi-speaker models by sharing data across languages and speakers. Luong et al. [8] shared speech data across speakers and built multi-speaker models from over 100 speakers by using speaker-code vectors and managed to control the produced voice's gender, age, or identity.

Nevertheless, building high-quality emotional speech synthesizers by using DNNs is a challenging topic. One reason is that DNN training typically require substantial amounts of speech data that cover various positive and negative emotions and their properly annotated labels. Moreover, annotating emotions itself is much more difficult and subjective compared with annotating reading speech, and the basic question of what kind of the supervised labels are necessary and important for training the DNN-based TTS systems has not been answered yet. Likewise, it is totally unknown how to construct DNN-based speech synthesizers that are capable of precisely controlling multiple emotional categories and emotional strength.

With that objective in mind, in this paper we investigate simultaneous modeling schemes of multiple emotions and how to represent emotional labels such as the emotional class. Our goal is to answer the question of what is the best way to annotate speech data with multiple emotions – should we use the labels that the speaker intended to express, or labels based on listener perception of the resulting speech signals? 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 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, annotated with perceived emotional labels from crowd-sourced listeners. By subjectively comparing synthetic speech generated from DNNs by using the different emotional representations, we assess the impact of these representations on human emotion recognition rates and perceived emotional strength.

The paper is structured as follows. First, we give a brief introduction to DNN-based speech synthesis in Section 2. Then, the proposed emotional representations are explained in Section 3. Section 4 introduces the emotional speech corpus used for building the DNN-TTS systems and also an explanation on the crowd-sourcing strategy that was applied to label the database. Then, the perceptual evaluation for the emotional synthetic speech is explained in Section 5. Finally, Section 5.4 shows the results of the perceptual evaluations. In Section 6, we draw global conclusions and discuss related work we expect to do in the future.

## 2. DNN-based speech synthesis

In the text-to-speech systems that use vocoders, extracted acoustic features are modeled by HMM-based or DNN-based acoustic models that represent the relationship between linguistic and speech features [1, 3, 9, 10]. In this section, we briefly review a DNN-based speech synthesis system [1].

Linguistic features obtained from a given text are mapped onto speech parameters by a DNN such as a feedforward neural network or recurrent neural network (RNN). The input linguistic features are composed of binary answers to questions about linguistic contexts and numeric values such as the number of words in the current phrase, the position of the current syllable in the word, and duration of the current phoneme. In [1], the output speech parameters include mel-cepstral coefficients and excitation parameters and their time derivatives (dynamic features). By using pairs of input and output features obtained from a training dataset, the parameters of a DNN can be trained with SGD [11]. Speech parameters can be predicted for an arbitrary text by a trained DNN using forward propagation.

## 3. Emotional representations for DNN-based speech synthesis systems

### 3.1. Discrete representations

#### *3.1.1. 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 as outlined in Figure 1. In our study, we use a RNN with long-short-term memory units (LSTM) [10].

If we have $C$-representative emotions as the talker categories, the one-hot vector $\boldsymbol{e}_i$ for the $i$-th emotion is defined as $\boldsymbol{e}_i = (e_1, e_2, ..., e_C)$, where each value $e_c$ is given by:

$$e_c = \begin{cases} 0 & (c \neq i), \\ 1 & (c = i). \end{cases} \quad (1)$$

Here, each subscript indicates the talker emotional category for a speech utterance. This simply 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.

#### *3.1.2. Representation based on listener dominant categories*

The above vector is "blind" because it does not represent the listener categories. It is indeed a fact that, even for professional speakers, the way we speak is not constant [12]. 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*.

Using the results of multiple listeners, we can re-label the emotional category of each sentence to a new class perceived by dominant listeners and may represent it by using a similar one-hot vector. Since the listeners may not perceive a talker's emotion as any of the $C$ emotions, we need to add an "other" emotional category into the one-hot vector.

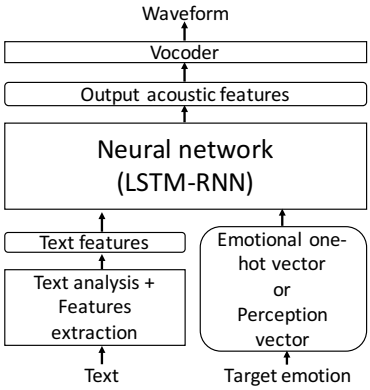


Figure 1: *Flowchart for the proposed RNN speech synthesis system using emotional one-hot or perception vectors.*

Table 1: *Confusion matrix of talker and listener emotional categories*

| | **Listener categories** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Talker's categories** | **1** | $\cdots$ | **j** | $\cdots$ | **C** | **O** |
| **1** | $e_{11}$ | | $e_{1j}$ | | $e_{1C}$ | $e_{1O}$ |
| **i** | $e_{i1}$ | | $e_{ij}$ | | $e_{iC}$ | $e_{iO}$ |
| **C** | $e_{C1}$ | | $e_{Cj}$ | | $e_{CC}$ | $e_{CO}$ |

### 3.2. Continuous representations

#### *3.2.1. 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 [13]. 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.

#### 3.2.2. 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 $\widehat{\boldsymbol{e}}_i = (e_{ij}, e_{i2}, ..., e_{iC}, e_{iO})$.

#### 3.2.3. 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. This is a natural extension of the one-hot vector based on listener categories. Note that since there is the "other" class, we need to normalize the values of $\overline{\boldsymbol{e}}_j$ as probabilities.

#### 3.2.4. Emotional confusion matrix based on listeners categories

In the above explanation, we have constructed the confusion matrix based on the basis of intended talker 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*.

#### 3.2.5. Units to compute the confusion matrix

The use of the continuous perception vectors as an additional auxiliary input to DNN-based acoustic models basically determines how we mix emotional speech data. If a global confusion matrix is used, we can obtain reliable statistics, but, speech data belonging to different emotional classes may be always mixed, and hence, reproduced emotional characteristics might be blurry compared with the original speech recordings. Hence, we may define a small subset including all the emotional categories as a mini-batch to be used for DNN training and may compute an emotional confusion matrix per the mini-batch.

An advantage of perception vector approaches is that DNNs can use more canonical data for each emotional category, but a disadvantage is the unbalanced quantity of speech data in each listener category, and it is required to have at least one sentence per talker emotional category to compute the confusion matrix.

### 3.3. 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 perceivable emotional strength. We may annotate the perceived emotional strength per sentence by computing the average across multiple listeners[1].

[1]It is also possible to compute the average strength per talker or listener emotional category per the mini-batch, but it would be more natural to use the sentence score per sentence given the fact that the fluctuations in emotional strength may happen sentence by sentence

Table 2: *Description of the Japanese emotional speech database. Audio duration includes silences at the beginning and end of the utterance and is expressed in minutes. Speaking rate excludes silences and is expressed in phones per second. Total duration and average speaking rate for the whole database are also shown. Phone alignment was obtained on basis of HMM-based forced alignment.*

| Emotion | #Sentences | Audio duration | Speaking rate |
|---|---|---|---|
| Neutral | 1200 | 147 min | 10.39 phones/sec |
| Happy | 1200 | 133 min | 10.90 phones/sec |
| Sad | 1200 | 158 min | 9.04 phones/sec |
| Calm | 1200 | 154 min | 9.05 phones/sec |
| Insecure | 1200 | 141 min | 9.88 phones/sec |
| Excited | 1200 | 136 min | 10.51 phones/sec |
| Angry | 1200 | 148 min | 9.26 phones/sec |
| Total | 8400 | 1017 min | 9.86 phones/sec |

Table 3: *Description of the Japanese emotional database recording sentences. Third column "common" indicates if the sentences were used for recordings of other emotional categories.*

| Source | Sentences | Common |
|---|---|---|
| News | 101 | Yes |
| Novel | 313 | No |
| TED talks | 196 | Yes |
| Car navigation system | 200 | Yes |
| MULTEXT | 191 | Yes |
| Phonetically balanced | 199 | Yes |
| Total | 1200 | |

## 4. Emotional speech corpus

### 4.1. Details of the emotional speech corpus

The emotional speech corpus used for this study is a self-recorded database consisting of three pairs of acted emotions uttered by a professional Japanese voice actress: happy - sad, calm - insecure, excited - angry in addition to neutral reading speech. A detailed description of the amounts of data per emotion can be seen in Table 2.

When recording the above emotional speech data in a studio booth, the voice actress was instructed to act out each emotion consistently (rather than changing emotional expressions depending on the meanings of sentences every time) in order to minimize variations within each emotion. This is a typical strategy used for speech synthesis.

The recorded sentences were chosen to have no emotional meaning, and hence may also be used for recordings of other emotional categories. Such sentences were carefully chosen from conversational text resources such as TED Talks or MULTEXT [14], rather than from news text resources. Conversational texts made it easier for the voice talent to express the emotions compared with news sentences. We also used novel sentences for the recording, but we manually filtered out sentences that induced emotional context emotion-by-emotion. Phonetically balanced sentences were also recorded to guarantee data availability for each phone. Please see Table 3 for a breakdown.

Table 4: *Confusion matrix of emotion recognition rates of the speech corpus in percentages. Rows show the talker's intended emotions and columns the perceived emotions. N stands for neutral, H for happiness, C for calm, E for excited, S for sad, I for insecure, A for angry and O for other.*

| | Listener categories | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Talker** | **N** | **H** | **C** | **E** | **S** | **I** | **A** | **O** |
| **N** | **78.6** | 0.7 | 4.9 | 0.6 | 0.3 | 1.0 | 4.6 | 9.3 |
| **H** | 1.3 | **84.7** | 2.6 | 6.0 | 0.2 | 0.3 | 0.1 | 4.7 |
| **C** | 18.3 | 2.5 | **71.5** | 1.5 | 0.9 | 1.3 | 0.1 | 4.0 |
| **E** | 1.2 | 30.4 | 1.3 | **32.7** | 0.2 | 0.2 | 5.0 | 29.1 |
| **S** | 0.3 | 0.7 | 0.2 | 0.0 | **81.7** | 14.1 | 0.4 | 2.5 |
| **I** | 0.7 | 0.0 | 0.9 | 0.1 | 24.2 | **71.7** | 0.1 | 3.0 |
| **A** | 0.7 | 0.2 | 0.2 | 0.6 | 0.0 | 0.6 | **91.0** | 6.7 |

### 4.2. Annotation of the emotional speech corpus

To obtain the perceived emotional categories and strength of the database, we designed and carried out a perceptual test. The evaluation was carried out by means of crowd-sourcing, where we asked Japanese natives of varied gender and ages to recognize the emotion of speech they were listening to and choose a strength value of the perceived emotion. For the emotional recognition question, they were asked to select an answer from a pool of nine emotions: neutral, happy, sad, excited, angry, calm, insecure, surprised, bored and "other". For the perceived emotional strength. they were asked to rank the strength on the MOS scale: from "1 - almost no emotion" to "5 - very emotional". They were also allowed to answer with "6 - no emotion". They were able to play the samples as many times as they wanted.

In total, we evaluated all sentences included in the emotional corpus twice. Each task consisted of a random selection of 14 sentences where subjects listened to every emotion twice, in random ordering. Evaluators were allowed to repeat the task up to 20 times to reduce the number of required evaluators, but they were not allowed to repeat the task too many times. Unless they completed the task completely, their results were deleted. In the end, a total of 266 listeners took part in the evaluation, for an average of 5 tasks completed per evaluator.

### 4.3. Confusion analysis of the emotional speech corpus

The obtained global emotional confusion matrix can be seen in Table 4. We can see that our acted emotional speech has confusion in terms of emotional categories as we expected. We can see that two thirds of the actress's "excited" emotion was incorrectly perceived as happy or other. In addition, we see that about 20% of her "calm" and "insecure" emotions were classed as neutral or sad, respectively.

### 4.4. Partitions of the database

The database was then fragmented into training, validation, and test sets consisting of an 80%, 10%, 10% of the data respectively. Both validation and test sentences were selected from those that both talkers and listeners agree 100% in terms of emotional categories, with an equal number of sentences per category. This makes comparisons of the vectors based on talker or listener categories fairer and our analysis easier. Note that our aim is to build TTS systems that have the least emotional confusion and to compare the systems' performance with natural speech that has 100% agreement. We also make sure that all the sentences in the validation and test have perceived emotional strength scores relatively close to the average. This helps us to have extremely strong or weak samples and make the test set a representative subset of the emotion.

Table 5: *Re-labeling strategy. Letters A to H represent two possible emotional categories, X represents an indifferent category, and O represents the "other" category.*

| **Talker** | **Listener 1** | **Listener 2** | **Re-labeled category** |
|---|---|---|---|
| B | A | A | A |
| B | B | X | B |
| B | X | B | B |
| B | C | A | O |

Table 6: *Confusion matrix after listeners category re-labeling. Description is analogous to Table 4.*

| | Listener perceived categories | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Emotion** | **N** | **H** | **C** | **E** | **S** | **I** | **A** | **O** |
| **N** | **85.5** | 1.0 | 8.5 | 0.0 | 0.2 | 0.5 | 0.4 | 3.9 |
| **H** | 0.3 | **82.8** | 1.2 | 7.4 | 0.6 | 0.0 | 0.0 | 7.7 |
| **C** | 5.0 | 2.1 | **88.0** | 0.4 | 0.2 | 1.6 | 0.2 | 2.5 |
| **E** | 1.2 | 10.7 | 1.9 | **81.1** | 0.2 | 0.0 | 1.2 | 3.7 |
| **S** | 0.2 | 0.2 | 0.5 | 0.0 | **83.3** | 13.9 | 0.0 | 2.0 |
| **I** | 0.8 | 0.5 | 0.7 | 0.2 | 12.6 | **82.1** | 0.5 | 2.4 |
| **A** | 3.1 | 0.1 | 0.1 | 1.6 | 0.3 | 0.1 | **90.5** | 4.2 |
| **O** | 14.9 | 6.5 | 1.9 | 14.2 | 2.2 | 4.5 | 7.7 | **48.0** |

### 4.5. Reflecting dominant perceived emotional categories

To take into account the information provided by the listeners, we also considered a re-labeling of the emotional categories of the database as we described earlier. For this re-labeling, we have used the following strategy, which is also adopted for US voting in general (See Table 5 for a visual guide):

- If both listeners agreed, the sentence was relabeled as being in the annotated category.
- If listeners disagreed, but one of them agreed with the intended talker category, it was left as in the talker category.
- If listeners disagreed, and no one agreed with the intended talker category, the sentence was relabeled as being in the "other" category.

The confusion matrix after this re-labeling process can be seen in Table 6. There, we can see how the re-labeling process brought the matrix closer to the identity matrix, although there was still a fluctuation due to disagreement between listeners. The "other" category appears as the representative of the uncertain category. A breakdown of the relabeled database can be seen in Table 7.

## 5. Experiment

The main objective of the experiment was to compare the modeling accuracy of DNN-based speech synthesizers by using the proposed emotional representations. The perceptual evaluation measured two aspects of emotional synthetic speech, that is, emotional strength, and emotion identification rates, although for this study, we only considered identification rates.

Table 7: *Distribution of the re-labeled database.*

| Emotion | #Sentences |
|---|---|
| Neutral | 1188 |
| Happy | 1281 |
| Calm | 1093 |
| Excited | 648 |
| Sad | 1219 |
| Insecure | 1128 |
| Angry | 1190 |
| Other | 628 |

### 5.1. RNN-based based neural-network system

The sampling rate of speech data was 48 kHz. From it we obtained a total of 259 acoustic features: 60 Mel-cepstral coefficients, linearly interpolated log $F_0$ extracted by using the SWIPE' algorithm [15], voiced/unvoiced parameter, and 25 band-limited aperiodicity coefficients. All the features were also represented by their $\Delta$ and $\Delta^2$ extracted with a temporal window of five frames.

389 Japanese linguistic features were obtained through the Open J-Talk engine [16]. We further appended the emotional vector and phoneme- and state-boundaries estimated by forced-alignment using 5-state left-to-right no-skip HMMs (using the HTS toolkit [17]). All the features besides the emotional vector were normalized to zero-mean unit-variance.

Both the one-hot vector and perception vector approaches were trained by using the same LSTM RNN-based architecture implemented within the CURRENNT toolkit [18]: two feed-forward layers and two bi-directional RNN LSTM layers with the layer size as (512, 512, 256, 256). Its output layer was a linear transformation layer. The activation function was sigmoid.

Training data was fed in mini-batches of 35 utterances, 5 sentences for each emotion to keep them balanced, for a total of 192 training mini-batches. The perception vector associated to each sentence was obtained from the confusion matrix of the mini-batch, and not of the whole database. The networks were randomly initialized and optimized by using SGD on the validation set, with early stopping after 40 epochs or on network convergence. The learning rate was fixed to $10^{-6}$.

The WORLD vocoder [19] was used for generating speech waveforms from the predicted acoustic features with the MLPG algorithm [20] for trajectory smoothing and also with a mel-cepstrum based post-filtering [21] with a 0.2 coefficient.

### 5.2. Evaluation

As we introduced in section 3, we can consider a number of emotion representations: based on listeners or talkers categories, based on the original labeling of the database or on the re-labeling, and with or without emotional strength information. We could also consider the one-hot vector representations, if only to serve as a reference.

In the present evaluation task, we considered the following 12 combinations of representations.

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)

Table 8: *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** | <u>0.63</u> | <u>1.41</u> |
| | **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** | <u>0.61</u> | <u>1.31</u> |
| | **Listener** | **One-hot** | 0.81 | 1.48 |
| | | **Talker** | 0.75 | 1.41 |
| | | **Listener** | 0.82 | 1.54 |

At synthesis time, we considered utterances from the test split of the database, so durations were obtained from forced alignment, and the perceptual vectors used for synthesis were the labels annotated by the database evaluators.

### 5.3. Evaluation design

The perceptual evaluation was also conducted by using crowdsourcing. The evaluators were asked to listen to a set of 14 emotional utterances one by one, without any additional information about the synthesized text or emotion. The samples could be played as many times as the listeners needed. Evaluators were asked to identify the emotion conveyed by the utterance from an open list of the synthesized emotions, including "neutral" and "other" options. Then, they were asked to rate the perceived emotional strength by using a five-point MOS scale. The utterances were presented to the listeners in random order without repetitions with the only constraint that every emotion should be present twice.

### 5.4. Results

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

#### *5.4.1. Modeling accuracy*

Since our aim is to build TTS systems that have the least emotional confusion and to compare the systems' performance with natural speech, 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 (shown in Table 8). The shorter the distance, the closer we are to representing natural speech. We can also compute the Frobenius distance to the identity matrix, thus measuring how far the proposed emotional synthesizers are from the ideal situation where there is no confusion.

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. We can also see that emotional strength information improved accuracy only when used together with the listener categories, but not so much in the other representations. Finally, we can see that the best emotional representation in terms of Frobenius distance overall was one based on the talker labeling, listener categories, and emotional strength, for a distance of 0.61, significantly lower than any of the representations apart from the analogous representation without emotional strength.

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 (two listeners per sentence) and by the unbalanced distribution of each emotional category after the relabeling.

## 6. Conclusions and Future Work

We have presented two novel concepts. We have carried out one of the first studies on the capability of DNN-based emotional speech synthesis with a one-hot vector approach to characterize each modeled emotion. We have also proposed a way of exploiting annotation information available for our training corpus in an attempt to enhance the expressive capability of the generated synthetic speech. In the end, the proposed systems were analyzed by means of a perceptual evaluation.

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 adding emotional strength as a separate input 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.

## 7. References


[1] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *Proceedings of ICASSP*, pp. 7962–7966, 2013.

[2] X. Wang, S. Takaki, and J. Yamagishi, "A comparative study of the performance of HMM, DNN, and RNN based speech synthesis systems trained on very large speaker-dependent corpora," in *9th Speech Synthesis Workshop (SSW9)*, 2016, pp. 125–128.

[3] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Interspeech*, 2014, pp. 1964–1968.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[5] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[6] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfiltering for statistical parametric speech synthesis," in *ICASSP*, 2017, pp. 4910–4914.

[7] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," *INTERSPEECH 2016*, pp. 2468–2472, 2016.

[8] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *ICASSP*, 2017, pp. 4905–4909.

[9] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 2129–2139, 2013.

[10] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," *Proceedings of Interspeech*, pp. 2268–2272, 2014.

[11] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science 28*, vol. 313, no. 5786, pp. 504–507, 2006.

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

[13] T. Bänziger, S. Patel, and K. R. Scherer, "The role of perceived voice and speech characteristics in vocal emotion communication," *Journal of nonverbal behavior*, vol. 38, no. 1, pp. 31–52, 2014.

[14] K. Shigeyoshi, K. Tatsuya, M. Kazuya, and I. Toshihiko, "Preliminary study of japanese MULTEXT: a prosodic corpus," in *International Conference on Speech Processing, Taejon, Korea*, 2001, pp. 825–828.

[15] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[16] K. Oura, S. Sako, and K. Tokuda, "Japanese text-to-speech synthesis system: Open JTalk," in *Proc. ASJ Spring*, 2010, pp. 343–344.

[17] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[18] F. Weninger, J. Bergmann, and B. Schuller, "Introducing currennt–the munich open-source cuda recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, no. 3, pp. 547–551, 2015.

[19] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[20] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[21] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *Systems and Computers in Japan*, vol. 36, no. 12, pp. 43–50, 2005. [Online]. Available: http://dx.doi.org/10.1002/scj.20354