Manuscript accepted for publication in Speech Communication. https://doi.org/10.1016/j.specom.2018.03.002

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

Jaime Lorenzo-Trueba[a,*], Gustav Eje Henter[a], Shinji Takaki[a], Junichi Yamagishi[a,b], Yosuke Morino[c], Yuta Ochiai[c]

[a]*National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan*
[b]*The University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, UK*
[c]*Toyota Motor Corporation, Higashifuji Technical Center, Shizuoka Prefecture 410-1193, Japan*

Abstract

In this paper, we investigate the 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 voice 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, perceived emotional strength, and subjective speech quality. Simultaneously, we also study which representations are most appropriate for controlling the emotional strength of synthetic speech.

Keywords: Emotional speech synthesis, perception modeling, perceptual evaluation

1. Introduction

Speech synthesis, or text-to-speech (TTS), is a long-studied technology that aims to generate natural-sounding, intelligible speech from arbitrary text. Currently, neural network (NN)-based TTS methods are being actively investigated because they are able to significantly improve the quality and naturalness of synthetic speech compared to traditional approaches (Zen et al., 2013; Watts et al., 2016). First, feed forward (FF) deep neural network (DNN) systems were proposed as a replacement for decision-tree approaches in HMM-based speech synthesis. These DNNs have demonstrated better synthesis accuracy from large amounts of speech data than traditional approaches (Wang et al., 2016). Long short-term memory (LSTM)-based recurrent neural networks (RNNs) have since been adopted, with reports that they provide even better naturalness and prosody of synthetic speech due to their capability to model the long-term dependencies of speech (Fan et al., 2014).

Lately, we have seen a sharp rise in the number of proposals of techniques based on waveform modeling (van den Oord et al., 2016; Mehri et al., 2016). 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 (Mehri et al., 2016). There have also been significant improvements in signal quality thanks to the implementation of generative adversarial networks (Kaneko et al., 2017). In this case one of their biggest keys to success is that they are capable of generating output through random sampling.

We have also seen a number of recent advances that demonstrate control capabilities for different voice aspects, such as (Li and Zen, 2016), who built multi-language and multi-speaker models by sharing data across languages and speakers. (Luong et al., 2017) built multi-speaker models from over 100 speakers using speaker-code vectors and managed to control the produced voice's gender, age, and identity.

Nevertheless, building high-quality emotional speech synthesizers by using DNNs is a challenging topic. One reason is that DNN training typically requires 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 schemes for the simultaneous modeling of multiple emotions

*Corresponding author
Email addresses: jaime@nii.ac.jp (Jaime Lorenzo-Trueba), gustav@nii.ac.jp (Gustav Eje Henter), takaki@nii.ac.jp (Shinji Takaki), jyamagis@nii.ac.jp (Junichi Yamagishi), yosuke_morino@mail.toyota.co.jp (Yosuke Morino), yuta_ochiai_aa@mail.toyota.co.jp (Yuta Ochiai)

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 – 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 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. Simultaneously, we also study which representations are most appropriate for controlling the emotional strength of synthetic speech.

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. By subjectively comparing synthetic speech generated from DNNs by using different emotional representations, we assess the impact of these representations on human emotion recognition rates, perceived emotional strength and subjective speech quality.

The paper is structured as follows. First we give an overview of emotional speech synthesis in Section 2 and on DNN-based speech synthesis in Section 3. Then, the proposed emotional representations are explained in Section 4 and the proposed emotion control technique in Section 5. Section 6 introduces the emotional speech corpus used for building the DNN-TTS systems and also explains the crowd-sourcing strategy that was applied to label the database. Then, the perceptual evaluation for the emotional synthetic speech is explained in Section 7. Finally, Section 8 shows the results of the perceptual evaluations. In Section 9, we draw some global conclusions and discuss work we expect to do in the future.

2. Overview of emotional speech synthesis

The main focus of speech synthesis is to achieve human-like levels of naturalness and intelligibility, which is slowly but steadily being achieved when talking about read speech. Human communication, though, consists of much more than just read speech: aspects such as conveying emotions, engaging in behaviour-driven dialogues, etc. All require the synthesis system to consider aspects such as speaking styles or emotional speech to achieve actual naturalness. A good sample can be seen in 2016 Blizzard Challenge (King and Karaiskos, 2016), where the task was to generate a voice based on a set of audiobook speech files. Results showed how the systems that did not pay special attention to the expressive nature of the data suffered considerably in the evaluation results.

In terms of speech synthesis technologies, all of the existing ones have shown some degree of success, each with their advantages and their disadvantages:

HMM-based approaches, thanks to their adaptation capabilities have been significantly successful in the task even for low amounts of training data (Yamagishi et al., 2005), and remain in use even nowadays for both high numbers of emotional classes (e.g. "the big six": anger, disgust, fear, happiness, sadness and surprise (Cabral et al., 2016)) and low number of emotional classes (e.g. emphasis vs. neutral (Do et al., 2016)). Their main advantage is the possibility of generating speaker-independent emotional speech models that aggregate information from different speakers and different emotions, providing highly controllable and reliable systems. Some tasks they have proved to been successful on range from emotional intensity control of 2 emotional classes (joyful and sad) (Nose and Kobayashi, 2012) to emotion transplantation of 4 emotional classes (angry, happy, sad and surprised) (Lorenzo-Trueba et al., 2015).

The main strength of unit selection-based approaches is the fact that they work with natural speech and hence are capable of providing extremely high quality and natural emotional speech when the selection of units and concatenations are successful, even for high number of emotional classes ("the big six") (Barra-Chicote et al., 2010; Erro et al., 2010). This main advantage is also their main weakness, as they are completely reliant on the corpus: it is not competitive to mix different speakers in order to obtain more robust systems, so there is a scalability problem when considering how difficult it is to obtain large amounts of emotional speech. Nevertheless, when properly designed and applied, unit selection systems (and nowadays hybrid unit selection systems) have proven to be very successful at the expressive speech synthesis task, especially for reduced numbers of emotional classes (e.g. affective and neutral) (Tsiakoulis et al., 2016).

A specific fact to expressive speech synthesis is that most of the ongoing work is still being done on HMM-based approaches, together with hybrid unit-selection based systems, with very little work being done yet in pure DNN-based systems, which the present paper aims to change. The latest work in the field of DNN-based expressive speech synthesis has shown that it is possible to replicate the HMM aggregation capabilities and then producing continuously controllable speech features such as age or gender (Luong et al., 2017), and some approaches have proven to be capable of continuously controlling the produced expressiveness (Watts et al., 2015).

An aspect of emotional speech synthesis that is still underdeveloped is that *emotional speech is not always perceived as it was intended when designing the corpus*. This information has been sometimes partially considered in commercial TTS systems (Pitrelli et al., 2006), but it is seldom done so in research because of the cost in obtaining perceptual annotations. As such, there is little research discussion of the topic and even less so concerning DNN-based systems. Nowadays, as shown by this research, the rise in crowd-sourcing platforms has made it feasible to reliably annotate large emotional speech corpora, allowing us to exploit this listeners' perceived annotations and compare their usefulness when compared with the traditionally

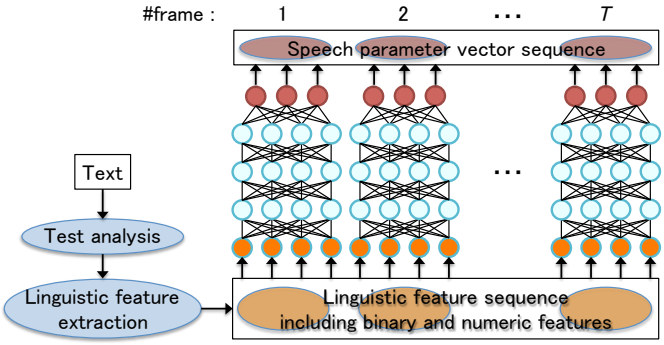

Figure 1: A framework for the DNN-based acoustic model.

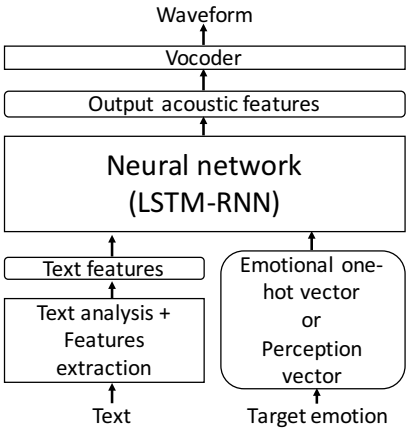

Figure 2: Flowchart for the proposed RNN speech synthesis system using the emotional one-hot or perception vectors.

used talkers' intended emotional categories.[1]

3. DNN-based speech synthesis

In the text-to-speech systems that use vocoders, the extracted acoustic features are modeled by HMM-based or DNN-based acoustic models that represent the relationship between linguistic and speech features (Zen et al., 2013; Fan et al., 2014; Ling et al., 2013; Fernandez et al., 2014). In this section, we briefly review a commonly known DNN-based speech synthesis system (Zen et al., 2013).

Figure 1 illustrates a framework of the DNN-based acoustic model where linguistic features obtained from a given text are mapped onto speech parameters by a DNN such as a feed-forward 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 durations of the current phoneme. In (Zen et al., 2013), 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 the DNN can be trained with SGD (Hinton and Salakhutdinov, 2006). Speech parameters can be predicted for an arbitrary text by a trained DNN with forward propagation.

4. Emotional representations for DNN-based speech synthesis systems

4.1. Discrete representations

4.1.1. 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 2. In our study, we use a RNN with long short-term memory units (LSTM) (Fernandez et al., 2014).

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.

4.1.2. 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 (Athanasopoulou and Vogel, 2016). 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 and have them annotate the emotional categories that they perceive when they listen to the emotional speech. We call the categories *listener emotional categories*.

Using the results of multiple listeners, we can re-label the emotional category of each sentence to a new class dominantly perceived by 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.

4.2. Continuous representations

4.2.1. Emotional confusion matrix based on talker categories

The above one-hot vectors are discrete representations of emotions, although it is claimed that the emotional space is a

[1] For a more in depth analysis on the history of expressive speech synthesis the authors suggest reading Marc Schröder's review (Schröder, 2009).

	Listener categories					
Talker 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}

Table 1: Confusion matrix of talker and listener emotional categories

continuum rather than a discrete space (Bänziger et al., 2014). 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 listeners categories. Here, e_{ij} shows the probability that the talker's i-th emotion is perceived as the j-th emotion by listeners.

4.2.2. 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 *talker categories weighted by listeners' perception*, which is a natural extension of the one-hot vector based on talker categories and is represented as $\widehat{\boldsymbol{e}}_i = (e_{i1}, e_{i2}, ..., e_{iC}, e_{iO})$.

4.2.3. 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, and a vector $\overline{\boldsymbol{e}}_j = (e_{1j}, e_{2j}, ..., e_{Cj})$ 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.

4.2.4. Emotional confusion matrix based on listeners categories

In the above explanation, we have constructed the confusion matrix 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*.

4.2.5. 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. As such, 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 the perception vector approaches is that DNNs can use more canonical data for each emotional category, but a disadvantage of the perception vector approaches 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.

4.3. 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[2].

5. Emotional control

The second objective of this research is to study which representations are most appropriate for controlling the emotions of synthetic speech. More specifically we want to be able to produce enhanced or de-enhanced versions of the learned emotions by modifying the emotional representation at synthesis time so that we can generate more versatile expressive speech.

5.1. Controlling the perception vectors

In the case of the perception vectors, we have two choices: modifying the emotional confusion matrix or strength value.

5.1.1. Modifying the emotional confusion matrix

For enhancing emotions based on the emotional confusion matrix, we may explicitly reduce the contributions of any confusable emotions and may make it more recognizable or stereotypical. More specifically, we can achieve this by using a *confusion-reduced* vector, that is, increasing the probability of the main emotion and reducing the probabilities of the other emotions. Each value e_c of the perception vectors may be modified as

$$e'_c = \begin{cases} e_c - \frac{\alpha}{C-1} & (c \neq i), \\ e_c + \alpha & (c = i). \end{cases} \quad (2)$$

where e_c represents the default value obtained from training data such as the average and α is an adjustable parameter. Alternatively we may make the probability of the main emotion to one and reducing the probabilities of the other emotions to zeros as an extreme strategy as follows:

$$e''_c = \begin{cases} 0.0 & (c \neq i), \\ 1.0 & (c = i). \end{cases} \quad (3)$$

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

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 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.

5.1.2. *Modifying the emotional strength*

We may also emphasize or de-emphasize the synthetic speech by manipulating emotional strength values. We may increase or decrease the default strength values obtained from the training data or may specify arbitrary emotional strength value scores. Any modifications are possible, but, for instance, we may add an adjustable parameter β to the default strength value obtained from training data such as the average. Alternatively we may set the maximum value used in listening tests (such as 5 in the 5-point MOS test) to the strength score.

5.2. *Bounding the modification rages*

The above manipulations using extreme values may result in significantly lower quality of synthetic speech. It is therefore reasonable to bound the modification ranges based on the training data. One possible way is bound the range as $\mu_i - K\sigma_i < s < \mu_i + K\sigma_i$ where i represents talker's or listener's i emotional category, s represents the emotional strength value that we want to use and μ_i and σ_i show the mean and standard deviation in the talker's or listener's i emotional category. K is typically 2 or 3. Likewise we can also accumulate statistics of e_c per the mini batch and bound e_c using the mean and standard deviation.

6. Emotional speech corpus

6.1. *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.

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	

Table 3: Description of the Japanese emotional database recording sentences. The third 'common' column indicates if the sentences were used for recordings of other emotional categories.

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 (Shigeyoshi et al., 2001), 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 sentences from novels 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.

6.2. *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 order. Evaluators were allowed to complete up to 20 tasks to reduce the number of required evaluators, but they were not allowed to complete too many tasks. Only results from completed tasks were included in the analysis. In the end, a total of 266 listeners took part in the evaluation, for an average of 5 tasks completed per evaluator. Age and gender distributions of the evaluators are shown in Table 4.

6.3. *Confusion analysis of the emotional speech corpus*

The obtained global emotional confusion matrix can be seen in Table 5. We can see that our acted emotional speech has confusion in terms of emotional categories as we expected. We

Age	Count
18-29	35
30-39	77
40-49	81
50-59	74
60+	15

Gender	Count
Female	134
Male	132

Table 4: Age and gender of the evaluators.

	Listener perceived 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

Table 5: 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.

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.

The histogram of the perceived emotional strength scores is shown in Figure 3. Neutral is excluded as it is not clearly identifiable what emotional strength means for the non-emotional case. We can see how all emotions present different emotional strength profiles. Most notably we can see how the *very low* and *low* emotional strength categories are almost never considered for all emotions, and also how the *very high* category is significant for both angry and sad voices. As a summary, we can say that, in this corpus, happy, calm, excited and insecure voices show an average medium-high emotional strength and that sad and angry voices are on average highly emotional.

6.4. 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 exclude extremely strong or weak samples and make the test set a representative subset of the emotion.

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

Re-labeled	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**

Table 7: Confusion matrix after listeners category re-labeling. Description is analogous to Table 5.

6.5. 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 process, when there was a disagreement between the labels we solved it according to the following protocol (See Table 6 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 7. 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 8.

7. Experiments

The main objective of the experiments is to compare the modeling accuracy and controllability of DNN-based speech synthesizers using the proposed emotional representations. The perceptual evaluation measured mainly two aspects of emotional synthetic speech: emotional strength, and emotion identification rates.

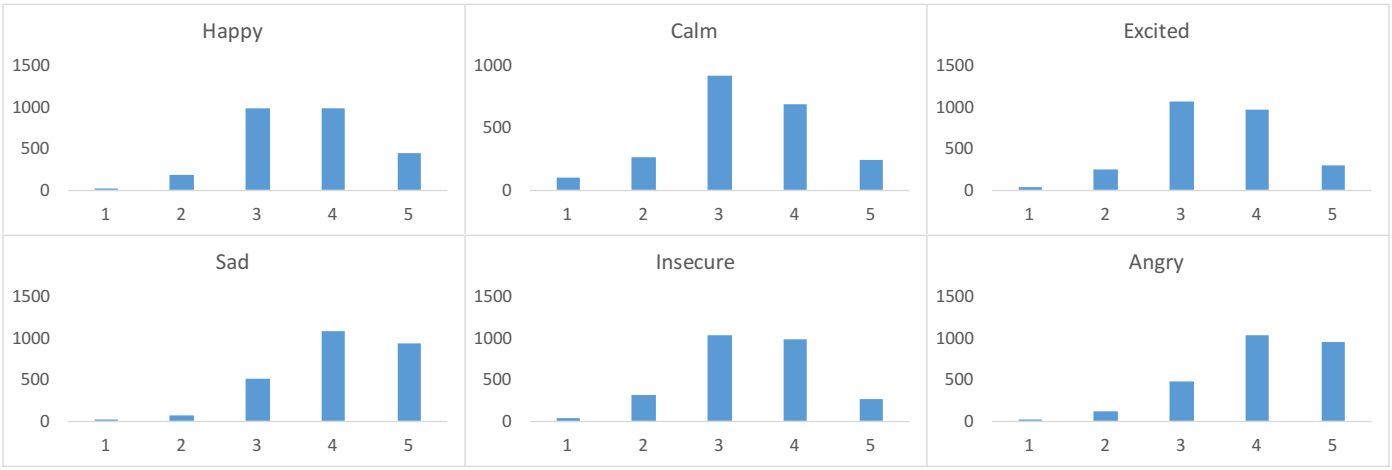

Figure 3: Histograms of the labelled perceived emotional strength of the database. Horizontal axis shows emotional strength from 1 "very low" to 5 "very high". Vertical axis shows the number of utterances associated to that emotional strength category.

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

Table 8: Distribution of the re-labeled database.

7.1. RNN-based based neural-network system

Speech analysis: The sampling rate of the speech data was 48 kHz. The speech data was first analyzed using pitch-adaptive spectral analysis method called "CheapTrick" implemented in the WORLD toolkit (Morise, 2015) with 4096 FFT points and a frame shift of 5 ms. From results of the spectral analysis, we obtained a total of 259 low-dimensional acoustic features: 60 Mel-cepstral coefficients (obtained via the all-pass filter based frequency warping of 0.77 to approximate the Bark scale), linearly interpolated log F_0 extracted using the SWIPE' algorithm (Camacho and Harris, 2008), voiced/unvoiced parameter, and 25 band-limited aperiodicity coefficients based on Bark's critical bands (Zwicker, 1961). All the features were also represented by their Δ and Δ^2 extracted with a temporal window of 5 frames.

Text analysis: The 389 Japanese linguistic features were obtained thorough the Open JTalk engine (Oura et al., 2010). 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 (Zen et al., 2007)). All the features besides the emotional vector were normalized to zero-mean unit-variance.

Network architecture: Both the one-hot vector and perception vector approaches were trained using the same LSTM RNN-based architecture implemented within the CURRENNT toolkit (Weninger et al., 2015): two feed-forward layers and two bi-directional RNN LSTM layers with the layer size as (512, 512, 256, 256). The output layer was a linear transformation layer. The activation function was sigmoid.

Network training: 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 using SGD on the validation set, with early stopping after 40 epochs or on network convergence, learning rate was fixed to 10^{-6}.

Waveform generation: Based on the trained LSTM RNNs, the same type of acoustic features was predicted from texts included in the test set. Further WORLD vocoder (Morise et al., 2016) was used for generating speech waveforms from the predicted acoustic features, with the MLPG algorithm (Tokuda et al., 2000) for trajectory smoothing and also with a mel-cepstrum based post-filtering (Yoshimura et al., 2005) with coefficient 0.2.

7.2. Evaluation points

For the present evaluation we considered the following two tasks:

1. Comparison between the representation systems to find the best emotional representation
2. Manipulation of the perceptual vector to analyze the control capabilities of the best emotional representation

The first task allows us to find out which of the proposed representation systems presents the best emotional representation capabilities, mimicking more closely natural speech. The second task, on the other hand, aims to see if by manipulation the perception vector at synthesis time we can control the expressiveness of the system, be it by enhancing or de-enhancing the emotional capabilities.

7.3. Emotional representation evaluation

As we introduced in section 4, we can consider a number of emotion representations: based on listeners or talkers categories (see section 4.2), based on the original labeling of the database or on the listener annotations-based re-labeling (see section 6.5), and with or without emotional strength (ES) information. We can also consider the discrete one-hot vector representations (see section 4.1.1), if only to serve as a reference.

In the present evaluation we considered the models trained by using the following 12 combinations of representations:

1. Original labeling, one-hot vector (w. and w/o. ES)
2. Original labeling, talkers categories based on the rows of the confusion matrix (w. and w/o. ES)
3. Original labeling, listeners categories based on the columns of the confusion matrix (w. and w/o. ES)
4. Re-labeling, one-hot vector (w. and w/o. ES)
5. Re-labeling, talkers categories based on the rows of the confusion matrix (w. and w/o. ES)
6. Re-labeling, listeners categories based on the columns of the confusion matrix (w. and w/o. ES)

The evaluation thus aimed to assess the emotional representation capabilities of the 12 different trained acoustic models. At synthesis time we considered the utterances from the test split with their corresponding labels obtained during the annotation of the corpus, without applying any manipulation.

7.4. Emotional control evaluation

For the second evaluation we wanted to measure how much we can stretch the emotional capabilities of our best emotional representation, fixing the evaluated acoustic model and this time synthesizing with manipulated perception vectors. For the fixed acoustic model, we selected among the representations in section 7.3, choosing the one that showed the best accuracy when compared with natural speech.

In terms of the evaluation itself, we considered the perception vector manipulation strategy based on the database annotations (see section 5.1).

7.4.1. Manipulation of the perception vector

In order to measure the capabilities of the system we modified the confusion matrix with a number of α values obtained from the variance of the confusion matrices of the mini-batches of the training data. To cover a broad spectrum of generation perception vectors we evaluated the cases when $\alpha = -5\sigma, -3\sigma, -\sigma, 0, \sigma$. Some sample generation matrices can be seen in table 9 and table 10. It must be noted that the $\alpha = \sigma$ condition already reaches the 100% condition, so we did not need to test $\alpha = 3\sigma$ or 5σ.

To manipulate the emotional strength values we took the average emotional strength of the target category and modified it with a $\pm\beta$ of 3σ. This allowed us to cover both the normal expressiveness of the system and a enhanced and de-enhanced version.

This resulted in a total of 15 perception vector manipulation strategies to evaluate.

	Listener perceived categories							
Emotion	**N**	**H**	**C**	**E**	**S**	**I**	**A**	**ES**
N	**57.1**	5.6	9.3	5.6	5.5	5.5	5.5	1.8
H	9.3	**26.5**	9.3	17.4	9.3	9.3	9.3	3.2
C	6.4	6.5	**55.2**	5.6	6.3	6.4	6.3	2.1
E	6.5	6.9	6.5	**53.9**	6.5	6.5	6.5	2.2
S	8.0	8.0	8.0	8.0	**38.4**	13.5	8.0	2.6
I	7.7	7.7	7.7	7.7	10.2	**43.4**	7.7	2.1
A	2.8	2.6	2.6	3.3	2.7	2.7	**80.6**	3.1

Table 9: Manipulation of the confusion matrix for $\alpha = -5\sigma$ and $\beta = -3\sigma$. Description is analogous to table 5, with the addition of the emotional strength (ES) column and removing the other (O) column.

	Listener perceived categories							
Emotion	**N**	**H**	**C**	**E**	**S**	**I**	**A**	**ES**
N	**88.3**	1.2	4.9	1.1	1.1	1.1	1.1	2.4
H	1.9	**78.9**	1.9	9.9	1.9	1.9	1.9	5.0
C	1.3	1.4	**90.8**	1.4	1.3	1.3	1.3	4.9
E	1.4	1.7	1.3	**90.5**	1.3	1.3	1.3	4.7
S	1.6	1.6	1.6	1.6	**83.3**	7.1	1.6	5.0
I	1.5	1.5	1.6	1.5	4.1	**86.6**	1.5	4.9
A	0.7	0.5	0.5	1.1	0.5	0.5	**95.5**	5.0

Table 10: Manipulation of the confusion matrix for $\alpha = -\sigma$ and $\beta = +3\sigma$. Description is analogous to table 9

7.5. Evaluations design

The perceptual evaluation was also conducted by using crowd-sourcing. 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. In the emotional control evaluation they were also asked to rate the perceived speech quality using a five-points MOS scale. The utterances were presented to the listeners in random order without repetitions with the only constraint that every emotion should be presented twice per set.

8. Results

A total of 54 native Japanese speakers took part in the first evaluation, and 102 in the second, for a total of 18270 utterance evaluations.

8.1. Emotion Identification

8.1.1. Impact of emotion representation manipulation at training time

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

Inputs	Labeling	Categories	Vs. Nat	Vs. ID
Emotional category	**Original**	**One-hot**	0.89	1.53
		Talker	0.70	1.49
		Listener	0.63	1.41
	Re-label	**One-hot**	0.95	1.68
		Talker	0.74	1.42
		Listener	0.75	1.50
Emotional category and strength	**Original**	**One-hot**	0.95	1.59
		Talker	0.75	1.40
		Listener	0.61	1.31
	Re-label	**One-hot**	0.81	1.48
		Talker	0.75	1.41
		Listener	0.82	1.54

Table 11: 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.

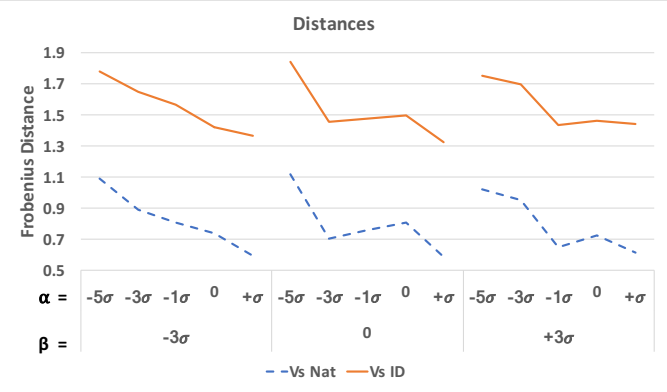

Figure 4: Distances of the evaluated manipulation configurations to the considered confusion matrices averaged across all emotions. The three different lines correspond to the three emotional strength manipulation values (β), and the lines represent the effect of the manipulation of the perception (α).

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 11). 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 the 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 annotations-based re-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 original 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 re-labeling.

8.1.2. Impact of emotion manipulation at synthesis time

In the emotion control case, as explained in section 7.4, we only considered the most accurate system (talkers labeling, listeners categories with ES) in order to get a reasonable number of systems for the evaluation.

In terms of the impact of the manipulation of the perception vector at synthesis time in accuracy, we can see in Figure 4 how there is a clear contraction in the distance the higher the α, reaching the minimum distance values for $\alpha = +\sigma$, which corresponds to synthesizing with the target emotion without contributions from other emotions. This happens to both the distance to the confusion matrix for natural speech and to the identity matrix, which means that we are both improving the absolute emotion identification rates, that is, distance to the identity matrix, and also improving the modeling accuracy at synthesis time (i.e. distance to the natural speech confusion matrix).

The impact of manipulating β, on the other hand, show an impact on the shape of the distances of their respective α manipulation, rather than on the absolute value themselves, with $\beta = -3\sigma$ alone providing monotonously decreasing results. This means that while adding the emotional strength component to the perception vector had a significant impact in modeling accuracy for the modeling step, it does not have a significant impact at the generation step.

In the end, it must be said that while at generation time we obtain the best accuracy results when synthesizing with the target emotion without adding any other emotion's information, this is improving over the results we achieved thanks to modeling the emotional data with the considered strategy. The effects of synthesizing with only the target emotion in the traditional one-hot vector modeling approach can be seen in Table 11 and are considerably worse than what we achieve in this configuration.

8.2. Emotional strength and speech quality

When considering emotion controllability, the objective is to be able to manipulate the perceived emotional strength generated at synthesis time, representing our capability to produce emphasized or de-emphasized emotional speech on demand. Figure 5 shows the results of our evaluations for the different combinations of α and β manipulations.

It is once again clear how α significantly impacts the perceived ES, with higher α values producing more emotional speech. This reinforces the idea that α manipulation is a strong way of manipulating the synthesized emotional speech. Again, β manipulation basically affects the result's shape, showing no significant impact in the absolute perceived ES results.

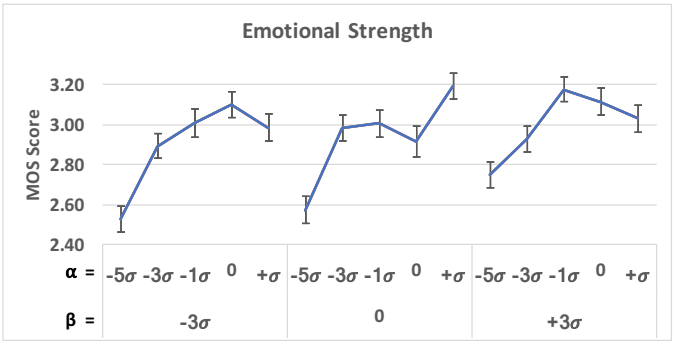

Figure 5: Evaluation results for emotional strength averaged across all emotions. The three different lines correspond to the three emotional strength manipulation values (β), and the lines represent the effect of the manipulation of the perception (α). Error bars represent 95% confidence intervals.

We also evaluated perceived speech quality, but we do not provide any figure as the results are not too interesting: we observed a non-significant variation of at maximum 0.05 points in MOS scale of perceived SQ reduction between $\alpha = -5\sigma$ and $\alpha = +1\sigma$. This means that while a less emotional configuration provides perceived SQ values closer to neutral and natural speech, the degradation introduced by our manipulation system is minimal and not significantly effective on the overall quality of the system.

To summarize, modifying the values in the confusion matrix has shown to be more suitable for controlling emotional speech than manipulating the emotional strength value itself.

9. Conclusions and Future Work

In this paper we present a number of novel contributions to the expressive speech synthesis field. First of all we have introduced a large scale analysis of an emotional speech corpus based on crowdsourcing. This has allowed us to present our second novel contribution, an analysis of different modeling strategies in an attempt to exploit perceptual information for improving the modeling accuracy of our acoustic models. The third and final contribution is the proposition of a mathematically-founded emotion control approach that can be integrated with the proposed improvements on modeling accuracy.

The proposed systems were evaluated by means of crowdsourced perceptual evaluations with a total of 156 different evaluators. The evaluation showed how adding perceptual information to the acoustic modeling process significantly impacts modeling accuracy, allowing us to provide synthetic speech with emotional recognition rates closer to both natural speech and to the ideal 0% confusion system. Also, we have shown the effectiveness of the proposed emotion manipulation system, as it allows us to systematically control the produced emotional strength without introducing significant speech quality degradation.

In conclusion, and to answer the questions placed in the abstract, the evaluations 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 original talker labels with listener categories. On the other hand, we have also seen how in terms of controlling the produced emotional speech it is much more effective to manipulate the confusion matrix components than modifying the emotional strength component in the perception vector.

As future work we would like to expand our emotional speech corpus, to work with expressive speech other than emotional speech and to experiment with sub-sentence level emotional annotations to attempt improve the modeling accuracy of our systems.

Athanasopoulou, A., Vogel, I., 2016. Acquisition of prosody: The role of variability. Speech Prosody 2016, 716–720.

Bänziger, T., Patel, S., Scherer, K. R., 2014. The role of perceived voice and speech characteristics in vocal emotion communication. Journal of nonverbal behavior 38 (1), 31–52.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., Macias-Guarasa, J., 2010. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Communication 52 (5), 394–404.

Cabral, J. P., Saam, C., Vanmassenhove, E., Bradley, S., Haider, F., 2016. The ADAPT entry to the Blizzard Challenge 2016. In: Proceedings of the Blizzard Challenge 2016.

Camacho, A., Harris, J. G., 2008. A sawtooth waveform inspired pitch estimator for speech and music. The Journal of the Acoustical Society of America 124 (3), 1638–1652.

Do, Q. T., Toda, T., Neubig, G., Sakti, S., Nakamura, S., 2016. A hybrid system for continuous word-level emphasis modeling based on HMM state clustering and adaptive training. Interspeech 2016, 3196–3200.

Erro, D., Navas, E., Herndez, I., Saratxaga, I., 2010. Emotion conversion based on prosodic unit selection. Audio, Speech, and Language Processing, IEEE Transactions on 18 (5), 974–983.

Fan, Y., Qian, Y., Xie, F.-L., Soong, F. K., 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks. In: Interspeech. pp. 1964–1968.

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

Hinton, G. E., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 28 313 (5786), 504–507.

Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., Kashino, K., 2017. Generative adversarial network-based postfiltering for statistical parametric speech synthesis. In: ICASSP. pp. 4910–4914.

King, S., Karaiskos, V., 2016. Blizzard Challenge 2016. Http://www.festvox.org/blizzard/.

Li, B., Zen, H., 2016. Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis. INTERSPEECH 2016, 2468–2472.

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

Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., Montero, J. M., 2015. Emotion transplantation through adaptation in HMM-based speech synthesis. Computer Speech & Language.

Luong, H.-T., Takaki, S., Henter, G. E., Yamagishi, J., 2017. Adapting and controlling DNN-based speech synthesis using input codes. In: ICASSP. pp. 4905–4909.

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

Morise, M., 2015. Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. Speech Communication 67, 1–7.

Morise, M., Yokomori, F., Ozawa, K., 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE transactions on information and systems E99-D (7), 1877–1884.

Nose, T., Kobayashi, T., 2012. An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. Speech Communication.

Oura, K., Sako, S., Tokuda, K., 2010. Japanese text-to-speech synthesis system: Open JTalk. In: Proc. ASJ Spring. pp. 343–344.

Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., Picheny, M. A., 2006. The IBM expressive text-to-speech synthesis system for American English. Audio, Speech, and Language Processing, IEEE Transactions on 14 (4), 1099–1108.

Schröder, M., 2009. Expressive speech synthesis: Past, present, and possible futures. In: Affective information processing. Springer, pp. 111–126.

Shigeyoshi, K., Tatsuya, K., Kazuya, M., Toshihiko, I., 2001. Preliminary study of japanese MULTEXT: a prosodic corpus. In: International Conference on Speech Processing, Taejon, Korea. pp. 825–828.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. 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, pp. 1315–1318.

Tsiakoulis, P., Raptis, S., Karabetsos, S., Chalamandaris, A., 2016. Affective word ratings for concatenative text-to-speech synthesis. In: Proceedings of the 20th Pan-Hellenic Conference on Informatics. ACM, p. 75.

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

Wang, X., Takaki, S., Yamagishi, J., 2016. 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). pp. 125–128.

Watts, O., Henter, G. E., Merritt, T., Wu, Z., King, S., March 2016. From HMMs to DNNs: where do the improvements come from? In: Proceedings of ICASSP. Vol. 41. pp. 5505–5509.

Watts, O., Wu, Z., King, S., 2015. Sentence-level control vectors for deep neural network speech synthesis. In: Proc. Interspeech. pp. 2217–2221.

Weninger, F., Bergmann, J., Schuller, B., 2015. Introducing currennt–the munich open-source cuda recurrent neural network toolkit. Journal of Machine Learning Research 16 (3), 547–551.

Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., 2005. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. IEICE TRANSACTIONS on Information and Systems 88 (3), 502–509.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2005. Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis. Systems and Computers in Japan 36 (12), 43–50.
URL http://dx.doi.org/10.1002/scj.20354

Zen, H., Senior, A., Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks. Proceedings of ICASSP, 7962–7966.

Zen, H., Toda, T., Nakamura, M., Tokuda, K., Jan. 2007. Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. Inf. & Syst. E90-D (1), 325–333.

Zwicker, E., 1961. Subdivision of the audible frequency range into critical bands (frequenzgruppen). The Journal of the Acoustical Society of America 33 (2), 248–248.