INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Direct Expressive Voice Training Based on Semantic Selection

Igor Jauk, Antonio Bonafonte

Universitat Politècnica de Catalunya
Barcelona, Spain
{igor.jauk, antonio.bonafonte}@upc.edu

Abstract

This work aims at creating expressive voices from audiobooks using semantic selection. First, for each utterance of the audiobook an acoustic feature vector is extracted, including iVectors built on MFCC and on F0 basis. Then, the transcription is projected into a semantic vector space. A seed utterance is projected to the semantic vector space and the N nearest neighbors are selected. The selection is then filtered by selecting only acoustically similar data.

The proposed technique can be used to train emotional voices by using emotional keywords or phrases as seeds, obtaining training data semantically similar to the seed. It can also be used to read larger texts in an expressive manner, creating specific voices for each sentence. That later application is compared to a DNN predictor, which predicts acoustic features from semantic features. The selected data is used to adapt statistical speech synthesis models. The performance of the technique is analyzed objectively and in a perceptive experiment. In the first part of the experiment, subjects clearly show preference for particular expressive voices to synthesize semantically expressive utterances. In the second part, the proposed method is shown to achieve similar or better performance than the DNN based prediction.

Index Terms: expressive speech synthesis, statistical speech synthesis, linguistic vector models

1. Introduction

Expressive and emotional information codified in text is definitely of interest for text analysis, and also for speech technology applications. In speech synthesis applications there might be two reasons why to *classify* text by expressiveness. One reason is to classify databases, such as audiobooks, to train synthetic voices. The other reason is to classify certain input text that could be "interpreted" by the machine. This work proposes a direct method of text classification for speech synthesis based on semantic selection. The proposed method not only can be used for input classification, but also for an *ad-hoc* voice generation from an expressive speech database, such as audiobooks.

There are several works that attempt predict emotions from text, such as [1, 2], using bag-of-words representations, knowledge and corpus based methods. Some studies have combined linguistic and acoustic features for emotion prediction form text, like [3, 4], where keywords and prosodic features are used in a call-center context in order to predict client's emotions. In [5] the authors bag-of-words models are combined with acoustic features to predict basic emotions; in [6] bag-of-words representations are mapped to continuous emotion representations an a three-dimensional space, as proposed by [7].

In speech synthesis, there have also been several successful attempts to use linguistic information for database clustering or input classification. In [8, 9] texts are being classified by domain and emotion using semantic representations. The classified data is used to train synthesizers. In [10, 11] emotions and expressiveness is predicted in form of *CAT* [12] model parameters that are used directly for expressive speech synthesis. In [11] expression and speaker modeling were separated from each other by training different sub-spaces for speakers and expressions.

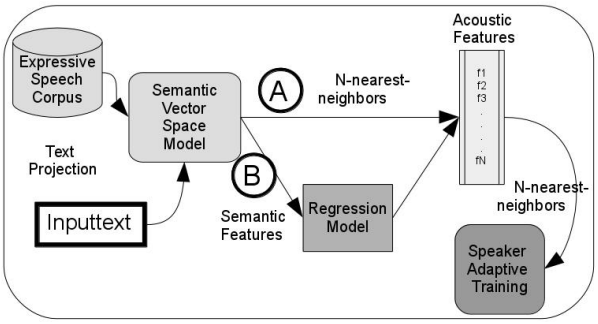

Figure 1: *Framework of the proposed approach.*

The present work proposes a novel expressive speech synthesis approach. Given a text excerpt to synthesize, each sentence is seen as an expressive *seed*. For each seed an appropriate expressive voice is created. In particular, statistical-parametric voice adaptation is performed using an appropriate part of an audiobook as database.

In works like [10, 11, 13] a prediction of acoustic features or parameters is performed from a semantic representation of a sentence. In this paper an alternative method is proposed, which uses directly a dual acoustic/semantic representation of an expressive corpus.

A particular application of the proposed method is the possibility to create emotional voices. Applying the method on a seed, sentence or word, which represents an emotion, an appropriate emotional voice is obtained, with no need of corpus labeling.

The rest of the article is structured the following way. Section 2 describes the general framework of the approach. Section 3 describes the databases and the linguistic and the acoustic features. Section 4 describes the experimental design, and section 5 present the perceptive results.

2. Framework

Figure 1 shows an overview of the proposed approach.

1. First, for each sentence of an expressive speech corpus, an audiobook in this case, an acoustic and a semantic

http://dx.doi.org/10.21437/Interspeech.2016-979

representations are created. The acoustic representation is a feature vector as explained in section 3.3. The semantic representation is a vector representation in a semantic vector space as explained in section 3.2. This dual representation of each utterance is base of the direct selection.

2. Then, an expressive *seed* sentence is projected into the semantic vector space, obtaining a vector representation of the seed.

3. **A).** *Direct selection*.

   - Semantic proximity.
     In the semantic vector space, N nearest neighbors to the seed are chosen from the expressive speech corpus.
   - Acoustic proximity.
     In the semantic vector space, only 1 nearest neighbors to the seed, the *seedling*, is chosen from the expressive speech corpus. Then, in the acoustic space N nearest neighbors to the seedling are chosen from the same expressive corpus.

4. **B).** *DNN prediction*.
   As a comparison framework, a Deep Neural Network, trained to predict acoustic features from the semantic features, predicts acoustic features for the seed. Then, in the acoustic space N nearest neighbors to the predicted vector are chosen from the expressive corpus.

5. The selected data is used to perform speaker adaptation in *HMM based speech synthesis*, using *HTS* tools [14] and the *AHOCoder* [15].

2.1. DNN predictor

A *Deep Neural Network* [16] was implemented to predict the feature vectors. The DNN is made of a stack of feed forward (*Dense*) layers, where each layer performs a projection followed by a nonlinearity, such that:

$$\mathbf{h} = g(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{1}$$

where $\mathbf{W}$ is the weights matrix, $\mathbf{x}$ is an input vector of features, $\mathbf{b}$ is the vector of biases and $\mathbf{g}$ is an element-wise non-linearity, which actually gives the DNN prediction capacity. There are several intermediate (hidden) layers, and in between, Dropouts [17] of 0.5 are applied to lower any possible over-fitting effect. At the output of the network a $\tanh$ activation function is used, so the output features are normalized between $[-1, 1]$.

Figure 2 shows the general architecture of DNNs used in this work. After several experiments the best network design turns out to be a bottleneck design. Since the entrance layer has a rather larger number of neurons (see section 3.2 and 4), the first hidden layer is also relatively large (1024 neurons). The next layer shrinks down to 256 neurons. There are several hidden layers with this number of neurons, which is then increased to 512, and to 620 in the output layer. Best results were achieved with 10 hidden layers.

3. Features

Since the framework of the study relies on an audiobook database, there is a set of conditions that should be fulfilled in order to optimally codify the linguistic and the acoustic representations. Eventually, the expressive synthesis is to be performed for large text instances, such as paragraphs. So the context of each sentence should be taken into account. The second point is that the audiobook in question, such as usual, contains many characters. And although all characters are being imitated by the same reader, the ways how they express themselves are very different. For example, *anger* will probably be expressed very differently by a giant than by a witch. So we need acoustic features that would represent the different characters as different speakers.

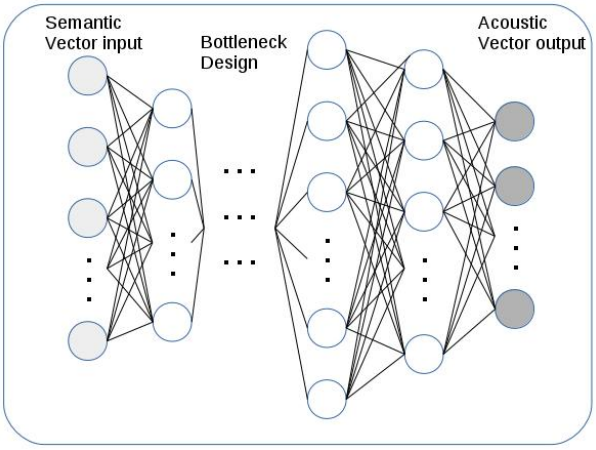

Figure 2: DNN architecture used for acoustic feature prediction.

3.1. Databases

The text database used for training of the semantic vector space model is the Spanish portion of the *Wikicorpus*, containing 120 million words [18].

The acoustic database is an audiobook of 8.8 hours of duration, segmented on sentence level and not labeled in no way. The topic is children or youth oriented. Some utterances that contained stuttering, reading errors, or noise imitations by the reader were removed, resulting in a total of 7903 utterances. The bad utterances have been identified partly by automatic tools and partly by manual revision. The segmentation was done using *Ogmios* speech analysis tools [19].

3.2. Linguistic Features

The linguistic features are coordinates of corpus sentences mapped into the *semantic vector space model (SVSM)*. The SVSM is trained using the *skip-gram* method [20] implemented in the *word2vec* package [21] with the *Wikicorpus*, resulting in a 600 dimensional vector space. The number of dimensions has been determined experimentally to provide best results under acceptable training and execution time conditions, though surely there is space to improvement. Each sentence is mapped into the space as a midpoint of the word embeddings of the sentence. One difference to most semantic vector space realizations is that in this work the function words have not been removed. The decision is inspired by studies presented in [22] and tested in previous studies (unpublished) on semantic representations with and without function words, where best results were achieved including the stop words.

3.3. Acoustic Features

Since the audiobook context implies not only the presence of different expressions, but also of different speakers, though only imitated by one reader, it is plausible to imply that acoustic features should account for the different speakers. A study conducted in [23] shows a significant performance improvement by including *iVectors* as a feature for unsupervised clustering of an audiobook. Also in [24] iVectors already have been used for emotion recognition.

iVectors represent speech in a low-dimensional total variability subspace, which leads to a representation that is independent of the different sources of variability such as speaker, channel or expressiveness, in this case.

First, acoustic features are extracted from the waveform; in this work, 40 Mel-frequency cepstral coefficients and F0 values are used. Before extracting the iVectors, a *Universal Background Model (UBM)* and the total variability matrix are trained as described in [25] and [26], respectively. In each case, the whole corpus was used for the training. The total variability matrix must be trained using audio segments that are homogeneous according to the speaker, channel and expressiveness. So silence was removed from the segments. Once the speech segments are obtained, Baum-Welch statistics are extracted using the UBM, which are used to obtain the total variability matrix that defines a total variability space, in which the speech segments are represented by a vector of total factors, namely iVector [27].

Traditionally, iVectors are calculated from MFCCs. Since prosody features are known to codify a significant amount of expressive information, in this work iVectors are also calculated from F0. Additionally syllable and silence rates, means, variance and medians of durations are added to the acoustic vectors. In result, the acoustic feature vectors are composed of 600 dimensional iVectors trained from MFCCs, 12 dimensional iVectors trained from F0, and 8 dimensional vectors with syllable and silence statistics, 620 dimensions in total. The MFCCs and F0 features were extracted using *AHOCoder* [15]. The syllable and silence duration using the *Ogmios* speech analysis tools [19], and the iVectors using the *Kaldi software* [28].

4. Experimental Design

Three experiments are conducted to test the viability of the proposed approach. The first task includes the reading of a paragraph of the book, which audiobook realization was used to train the voices. This task as a topline (proof of concept) since the audiobook is known. The paragraph was excluded from training of the DNN predictor. The paragraph is a dialogue between two book characters and narrator comments; it contains 16 sentences. Each sentence is projected into the SVSM and its coordinates are extracted. In the direct selection, 50 nearest neighbors are selected directly from the acoustic space and a voice is trained for each sentence (acoustic proximity). In the DNN prediction, acoustic coordinates are predicted for each sentence, which are then used as centroids to select 50 nearest neighbors in the acoustic feature space. Also, the paragraph is synthesized using a neutral voice, trained from approx. 10 hours of studio recorded read speech. The participants are asked to rank the three voices by best expressive performance and by quality.

The second task is similar to the first task, but the paragraph to synthesize is extracted from a new book that has not been projected into the SVSM nor used for the DNN training. Though, to maintain the context, this book is the continuation of the first book. Here, in the direct selection, for each input sentence 50 nearest neighbors are selected in the SVSM and used directly for voice training (semantic proximity). The DNN prediction is identical to the one in the first task: acoustic coordinates are predicted for each sentence, which are then used as centroids to select 50 nearest neighbors in the acoustic feature space. And also here, the paragraph is synthesized using a neutral voice.

Eventually, the semantic proximity is justified for this task because it provides higher semantic stability of the selected data, since the data is unknown. Using the acoustic proximity, there is a risk that the selected seedlings have not the appropriate expressiveness and the whole paragraph will be read with "jumping" expressive styles.

Since paragraphs are synthesized for the first two tasks, context can be used to achieve better prediction. So, each predictor vector is composed of the semantic vector for the sentence in question, and of two additional vectors, for the left and for the right contexts respectively. These context vectors are calculated using three closest words on the left and three closest words on the right of the sentence in question.

The third task evaluates the system's capability to create emotional voices. In this case the seeds are key words or key phrases that semantically represent the emotion of the desired voice. Three emotional voices have been trained: *happy*, from seed "happy", *angry*, from seed "I don't want!" and *suspense*, from seed "Mysterious secret in silent obscurity." For this task, the acoustic proximity is considered to be the best approach. However, the seedling was chosen manually by listening from a selection of 20 sentences in the SVSM. Then, 50 nearest neighbors to the seedling were selected in the acoustic space and used for voice adaptation. Although there is a manual selection involved, the effort is negligible in comparison to labeling of a corpus.

Using the three expressive voices and the neutral voice, seven sentences are synthesized. The sentences are designed to reflect semantically the emotion of the voices. The sentences are listed below, translated from Spanish.

- **Happy$_1$:** We have won the paella competition.
- **Happy$_2$:** Finally the holidays begin.
- **Angry$_1$:** You are an idiot. Never speak to me again.
- **Angry$_2$:** You are a goof-off. If you don't push yourself we won't win anything.
- **Suspense:** In the middle of the night, a silent shadow moved along the corridor.
- **Sad:** We haven't won the paella competition.
- **Neutral:** In many civilizations seven-days weeks are in use.

As seen in the sentence list, there is a *sad* sentence, but there is no sad voice. This is because we hypothesize that the *suspense* voice can also be used for *sad* or even *neutral* content. This might be also true for the *neutral* voice.

5. Perceptive Results

A total of 21 subjects participated in the experiment, some of them experienced with speech technology (either development or usage), and others not. Table 1 presents the results for the first two tasks. Participants had the option to prefer two synthesized paragraphs, or all of them, if sounded equally, again, regarding expressiveness and quality. In both tasks, regarding

expressiveness, the results show clear preference of both synthesis methods over the neutral voice. In the first task, or the proposed direct method was chosen to be better, or at least equal to the DNN based method. In the second task, almost half of the subjects chose the direct method and the DNN based method equally good. If not, the DNN method was slightly preferred.

In synthesis quality, there was no significant preference for none of the voices.

Table 1: Prediction method preferences by users for the first two tasks.

	DNN	direct	neutral	DNN =direct	direct =neutral
Task_1	0.19	0.43	0.0	0.38	0.0
Task_2	0.29	0.14	0.04	0.48	0.05

Table 2 shows the preferences for the third task. The first *happy* sentence is divided between the three expressive voices, with slight preference of the *angry* voice. For the second *happy* sentence there is a clear preference of the *happy* voice. A possible explanation for given distribution for the first *happy* sentence is that may be it is ambiguous to the listeners. However, the *happy* and the *angry* voices both sound similar and can be appropriate to both types of sentences. In fact, the *angry* voice does not sound really angry, it is more "book" angry, and meant for children.

For the first *angry* sentence there is a clear preference for the *angry* voice, with the *happy* voice on the second place. The second *angry* sentence is divided between the *happy* and the angry voice.

For the *suspense* sentence there is a very clear preference of the *suspense* voice. The *sad* and *neutral* sentences are also divided between the *suspense* and the *neutral* voice. There is no explicit *sad* voice, however the *suspense* voice does a good job imitating sadness, and as the results show, also the *neutral* voice.

Table 2: Task 3. Voice preference for each sentence.

	happy	angry	suspense	neutral
Happy_1	**0.29**	**0.38**	**0.24**	0.10
Happy_2	**0.52**	0.24	0.10	0.14
Angry_1	0.14	**0.48**	0.24	0.14
Angry_2	**0.38**	**0.43**	0.14	0.05
Suspense	0.0	0.05	**0.81**	0.14
Sadness	0.19	0.05	**0.43**	**0.43**
Neutral	0.10	0.05	**0.43**	**0.43**

6. Conclusions

In this work two methods were implemented to predict expressiveness from vector based text representation. The proposed method uses a direct selection method based on nearest neighbor selection in the semantic vector space, and then, refining the selection in the acoustic feature space. This method can also be used to create expressive voices from an unlabeled expressive corpus. In the second method a DNN predicts acoustic features from semantic vector coordinates.

Three experimental tasks were conducted in order to evaluate the proposed direct prediction method. Two paragraphs, one from a known, and one from an unknown book were synthesized. Both methods were compared between them and between a neutral voice. In the third task, semantically expressive sentences were synthesized using expressive voices trained with the proposed direct selection method. Subjects had to choose the best voice for each sentence.

The results of the paragraph synthesis show clear performance of both methods over the neutral voice. There is also a slight preference of the direct method for synthesis for a known book; and slight preference of the DNN based method for an unknown book.

The results of the third task show clear preferences of some voices over others for certain expressive sentences, that match the expectation. Nevertheless, there are interpretation differences of sentences, and also some voices seem to be useful for different expressions, such as the *suspense* voice can be used for *suspense*, *sad* and *neutral* sentences, and the neutral voices for *sad* and *neutral* sentences, but not for *suspense*.

The proposed method for expressive speech synthesis has proved to be useful for given applications. Although a particular voice is trained for each sentence, it is useful in offline applications. It offers high flexibility and a minimal manual work load. Future improvements can include controllability of the expressiveness, more stable semantic selection and improvement of synthesis quality.

In conclusion, the generated expressive voices have turned out to be much better than the neutral voice, and certainly much more vivid and appropriate for book reading. There have been a lot of positive comments about this point and basically the subjects had fun listening to the readings.

7. Acknowledgements

This work was supported by the Spanish Ministerio de Economía y Competitividad and European Regional Developmend Fund, contract TEC2015-69266-P (MINECO/FEDER, UE) and by the FPU grant (Formación de Profesorado Universitario) from the Spanish Ministry of Science and Innovation (MCINN) to Igor Jauk. A part of this work was realized during a research stay at the University of Texas at El Paso, also supported by the FPU grant.

8. References

[1] S. Ovesdotter Alm, D. Roth, and R. Sproat, "Emotion from text: machine learning for text-based emotion prediction," in *Proceedings of Conf. HLT-EMNLP*, 2005, pp. 579–586.

[2] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of 2008 ACM Symposium on Applied Computing*, 2008, pp. 1556–1560.

[3] D. L. and L. Lamel, "Emotion detection in task-oriented dialogues," in *Proceedings of ICME 2003*, vol. III, 2003, pp. 549–552.

[4] C. Lee and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proceedings of ICSLP 2002*, 2002.

[5] B. Schuller, R. Mller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Proceedings of Interspeech*, 2005, pp. 805–808.

[6] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space," *IEEE Transactions on affective computing*, vol. 2, no. 4, pp. 192–205, 2000.

[7] R. Kehrein, "The prosody of authentic emotions," in *Proceedings of Speech Prosody*, 2002, pp. 423–426.

[8] F. Alas Pujol, "Conversin de texto en habla multidominio basada en seleccin de unidades con ajuste subjetivo de pesos y marcado robusto de pitch," Ph.D. dissertation, Universitat Ramon Llull, 2006.

[9] O. Watts, "Unsupervised learning for text-to-speech synthesis," Ph.D. dissertation, University of Edinburgh, 2012.

[10] L. Chen, M. Gales, N. Braunschweiler, M. Akamine, and K. Knill, "Integrated expression prediction and speech synthesis from text," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 323–335, 2014.

[11] L. Chen, N. Braunschweiler, and M. Gales, "Speaker dependent expression predictor form text: Expressiveness and transplantation," in *Proceedings of ICASSP 2014*, 2014, pp. 2574–2578.

[12] L. Chen, M. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proceedings of Interspeech*, 2012, pp. 958–961.

[13] I. Jauk, A. Bonafonte, and S. Pascual, "Acoustic feature prediction from semantic features for expressive speech using deep neural networks," in *EUSIPCO 2016*, p. under review.

[14] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, "The HMM-based speech synthesis system (HTS)," 2008. [Online]. Available: http://hts.ics.nitech.ac.jp

[15] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Improved HNM-based vocoder for statistical synthesizers," in *Proceedings of Interspeech*, 2011, pp. 1809–1812.

[16] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] S. Reese, G. Boleda, L. Cuadros, M. Padr, and G. Rigau, "Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus," in *Proceedings of 7th Language Resources and Evaluation Conference (LREC10)*, 2010, pp. 1418–1421.

[19] T. Bonafonte, P. Aguero, J. Adell, J. Perez, and A. Moreno, "Ogmios: the UPC text-to-speech synthesis system for spoken translation," in *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, 2006, pp. 199–204.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at ICLR*, 2013.

[21] "word2vec - tool for computing continuous distributed representations of words," http://www.gnu.org/software/gsl/.

[22] J. Pennebaker, *The Secret Life of Pronouns*, 2011.

[23] I. Jauk, A. Bonafonte, P. Lpez-Otero, and L. Docio-Fernandez, "Creating expressive synthetic voices by unsupervised clustering of audiobooks," in *Proceedings of Interspeech 2015*, pp. 3380–3384.

[24] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "iVectors for continuous emotion recognition," in *Proceedings of Iberspeech 2014*, 2014, pp. 31–40.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models." *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[26] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data." *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[27] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.