9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Prosodic and Spectral iVectors for Expressive Speech Synthesis

Igor Jauk, Antonio Bonafonte

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

Abstract

This work presents a study on the suitability of prosodic and acoustic features, with a special focus on i-vectors, in expressive speech analysis and synthesis. For each utterance of two different databases, a laboratory recorded emotional acted speech, and an audiobook, several prosodic and acoustic features are extracted. Among them, i-vectors are built not only on the MFCC base, but also on F0, power and syllable durations. Then, unsupervised clustering is performed using different feature combinations. The resulting clusters are evaluated calculating cluster entropy for labeled portions of the databases. Additionally, synthetic voices are trained, applying speaker adaptive training, from the clusters built from the audiobook. The voices are evaluated in a perceptual test where the participants have to edit an audiobook paragraph using the synthetic voices.

The objective results suggest that i-vectors are very useful for the audiobook, where different speakers (book characters) are imitated. On the other hand, for the laboratory recordings, traditional prosodic features outperform i-vectors. Also, a closer analysis of the created clusters suggest that different speakers use different prosodic and acoustic means to convey emotions. The perceptual results suggest that the proposed i-vector based feature combinations can be used for audiobook clustering and voice training.

Index Terms: statistical speech synthesis, expressive speech, i-vectors

1. Introduction

The goal of the present paper is to study the usability of *i-vectors* for expressive speech synthesis, in comparison to more traditional features. i-vectors have been proved to be very useful in speaker verification applications (e.g. [1, 2]). Though, in recent works they have been proposed to identify emotional or expressive speech, as in [3, 4].

On the other hand, traditionally rather prosodic parameters have been used for expressive speech classification, synthesis etc. For instance, [5] uses glottal source parameters to perform clustering of expressive speech styles in audiobooks. In [6] a set of mainly prosody-based and some spectral based features is used for emotion recognition. In [7] prosodic features, i.e. F0, voicing probability, local jitter and shimmer, and *logarithmic HNR* are used for audiobook clustering and posterior synthetic voice training. As authors in [7] state, spectral features are considered to be poorly related to expressiveness. However, some approaches showed that spectral features are also important for the discrimination of expressiveness. Barra-Chicote et al. [8] suggest that different expressions are better characterized by different features; for instance, *anger* is rather characterized by spectral parameters, while *happiness* and *disgust* are better represented by both prosodic and spectral features.

Additionally, working with corpora such as audiobooks, or TV/radio programs, it has to be taken into account that there are several speakers present in the database, though in audiobooks the speakers (book characters) are usually imitated by the same reader, so it is an approximation to a multi-speaker database. In such corpora, generally not all speakers express all types of possible emotions (except maybe leading characters in a book), or they express them in different ways. For example, an angry giant would sound very differently than an angry hysterical witch. Concluding, there is need for features that are actually capable of not only account for emotions, but also for speakers.

As has been shown in [4], *i-vectors* have achieved the best results in audiobook clustering, in comparison to other features. However, only *MFCC* based i-vectors were used. In the present work, different i-vectors are trained on both, spectral and prosodic features, trying to achieve a balance between speakers and emotions in an audiobook clustering. Clustering results with i-vectors based features are compared to traditionally used features. The experiments are performed on two databases. An audiobook, including a large number of characters, i.e. imitated speakers; and an emotional database recorded by two speakers, where each imitates six basic emotions and a neutral voice. Additionally, synthetic voices are built on clusters created from the audiobook and are evaluated in a perceptual experiment.

The rest of the article is structured the following way: Section 2 describes the general approach. Section 3 describes the acoustic features used in the experiments. Section 4 describes the experimental design and the databases. Sections 5 and 6 present the results and finally section 7 draws final conclusions.

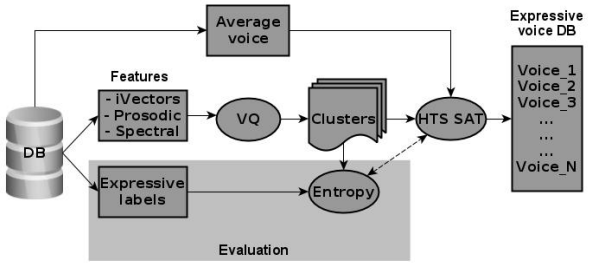
2. Framework

Figure 1 gives an overview of the system framework. First, a set of acoustic features, including i-vectors, is calculated for each utterance of an expressive database. Then, unsupervised clustering is performed using the extracted features and the resulting clusters are evaluated objectively, calculating cluster entropy. Two corpora are used for the evaluation. One is a laboratory recorded speech corpus, where each sentence has a label. The other one is the audiobook, where an excerpt was annotated manually providing speaker and expressiveness labels. These labels are used for the entropy calculations. More details on the corpora used for the experiments are given in section 4.1.

Further, synthetic voices are generated by means of HMM-based speech synthesis. To do so, an *average voice model (AVM)* is trained using HTS [9]. It must be noted that in this case the average voice model does not refer to a speaker independent voice model *(SI)*, since there is only one audiobook reader who tries to imitate different characters and expressions. Hence, *AVM* refers to the different characters and speaking styles imi-

http://dx.doi.org/10.21437/SSW.2016-10

Figure 1: System framework



tated by the audiobook reader.

Once the AVM is trained, different voices are adapted using the clustered speech segments. The voices are adapted using the Speaker Adaptive Training (SAT) technique [10], which performs a transformation of a model set λ applying speaker specific variations G^R, where R is the number of speakers. In this case, the speakers R are defined by the speech segments of each cluster, and they should represent different imitated voices and expressions produced by the reader of the audiobook. AHOCoder [11] is used to synthesize the voices.

3. Features

Working with emotional databases requires acoustic correlates, which account for the emotions and/or different speaking styles. Additionally, multi-speaker databases require features, which account for the different speakers, and for the different ways that speakers may express their emotions. I-vectors have proven to be very useful in speaker verification and classification tasks, and also in emotion or expressiveness classification, as in [4, 3]. On the other hand prosodic features generally have been considered useful for emotional speech analysis. In this work i-vectors built on prosodic features are proposed as an alternative or as an additional feature for expressive speech analysis. This following sections describe the features used in the proposed framework.

3.1. I-vectors

i-vectors represent speech in a total variability subspace, which leads to a representation that is independent of the different sources of variability such as speaker or channel.

First, acoustic and/or prosodic features are extracted from the waveform. In this work, i-vectors are calculated on:

- 40 Mel-frequency cepstral coefficients, extracted using the AHOCoder [11]. I-vectors dimension: 600
- Fundamental frequency of voiced segments, extracted using AHOCoder, where only the voiced parts are used for the i-vector extraction. I-vectors dimension: 12
- Power. I-vectors dimension: 16
- Syllable durations, calculated using forced alignment with *Ogmios* [12]. I-vectors dimension: 12

Before extracting the i-vectors, a *Universal Background Model (UBM)* and the total variability matrix are trained as described in [1] and [13], respectively. Both are trained using the whole database. The total variability matrix must be trained using audio segments that are homogeneous according to the speaker, channel and expressiveness. Therefore the training data is automatically divided into segments using a voice activity detector, eliminating silences. 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 i-vector [2].

3.2. Other Acoustic Features

Other acoustic features used for the experiments are prosody based. Among them are:

- F0 means, variance and range between the minimum and the maximum values. Extracted using AHOCoder.
- Syllable frequency and durations, means, variance and medians. Extracted from a forced alignment using Ogmios.
- Silence frequency and durations, means variance and medians. Extracted from a forced alignment using Ogmios.
- Local Jitter and Shimmer. Jitter is the period duration variation. Shimmer is the period amplitude variation. Extracted using *Praat* [14].
- Power.

Different feature set combinations are tested. Details are given in section 4.2.

4. Experimental Design

Two databases are used for the experiment, described in detail in section 4.1. For each sentence of both databases a feature vector is extracted. The resulting vectors are clustered applying a k-means algorithm, i.e. *vector quantization (VQ)* [15], performed using the *lbg* and *vq* tools from the SPTK toolkit [16].

The laboratory recorded database has a complete emotion annotation. The audiobook is only partly annotated with character and expression labels. Using the labels, the entropy is calculated for each cluster. Assuming that the clustering is successful, the resulting clusters would represent different expressions, speaking styles or characters. If that happens, then the information contained in each cluster would be rather monotonous. As defined in [17], and applied in clustering e.g. in [18]:

$$H(X_c) = -\sum_{i=1}^{q} \frac{n_c^i}{n_c} \log_2(\frac{n_c^i}{n_c}) \quad (1)$$

where X_c represents the whole set of speech segments in a cluster c, q is the number of labels (expressiveness or characters), n_c^i is the number of elements labeled with i occurring in cluster c and n_c is the number of elements assigned to the c^{th} cluster. Given a set X of labeled segments, if all the elements assigned to a cluster have the same label, the cluster would be fully homogeneous and its entropy would be 0. The other way around, if all elements in a cluster have different labels, then the cluster would be random and its entropy would be maximum. The best features are those that achieve minimum entropy.

The weighted entropy regarding the cluster size is computed as:

$$\bar{H}(X_c) = \frac{\sum_{c=1}^{C} H(X_c)\, n_c}{K} \quad (2)$$

where K is the number of all elements in the database, C is the number of clusters generated by VQ, $H(X_c)$ is the entropy of

the c^{th} cluster and n_c is the number of elements assigned to the c^{th} cluster.

Applying the clustering on the whole audiobook (not only the annotated part), voices were trained using *speaker adaptive training (SAT)* from the clusters created with the best feature set. The voices are then used in a perceptual experiment, where each participant has to edit a small paragraph of the audiobook. A total of two characters and the narrator appear in a dialogue of 18 sentences. The book scene is briefly described in order to introduce the characters and to provide some basic background knowledge to those participants who do not know the book or the topic. The participants have to choose between 10 voices and assign them to the characters. The voices are trained from clusters, which are acoustically closest to the original utterances. A total of 4 voice were chosen for the dialogue. The other 6 voices are assigned randomly from the other clusters. No examples are provided of how the real audiobook characters sound, so each participant can assign the voice that she considers to be most suitable for the characters.

4.1. Databases

Two databases are used in the experiment. The first one is a laboratory recorded emotional speech corpus in European Spanish, recorded by two professional speakers, male and female, with approx. 350 sentences each, recorded 7 times by each speaker, a total of 6.4 hours of duration [19]. The emotions recorded in the database are *angry*, *disgust*, *fear*, *joy*, *neutral*, *sadness* and *surprise*. Each sentence is recorded with each emotion. The female corpus will be referred to as C_1, the male corpus will be referred to as C_2.

The second database is an audiobook, with a total of 7900 sentences, and of 8.8 hours of duration. A part of the audiobook is labeled with expression and character (speaker) labels. The annotated part contains 1200 sentences, of a total duration of approx. 1.5 hours. Bad utterances have been identified partly by automatic tools and partly by manual revision. The annotated part does not contain *neutral* labeled speech. The labeled part of the audiobook will be referred to as A_l.

The expression or emotion labeling in the audiobook is not trivial, since emotions are expressed differently by different characters. There is a lot of scaling, i.e. different intensity of expressions, or combinations of emotions, such that for instance *surprise* can be negative, positive, or even finer-grained, it can be sad, joyful, aggressive, etc. So, the set of possible expressions is rather free, considering combinations of different expressive styles such as *surprise-anger* vs *surprise-joy*, also intending to label the intensity of the expressions. In total, a set of 248 different labels of expressiveness and 18 characters were obtained.

4.2. Feature sets

For the objective test different feature combinations are composed. Means, variance and medians are calculated for features F0, F0 range, syllable and silence durations, power, jitter and shimmer. The features are combined as follows:

- *Pitch*: F0 means, variance and range.
- *Rhythm*: Silence and syllable frequency and durations, means, variances and range.
- *JShimm*: Local jitter and shimmer.
- *iVecC*: F0 and MFCC based i-vectors.

The combinations are tested alone, and combined between them.

5. Objective Results

Table 1: Entropies for different features combinations and for the three databases. For the audiobook database (A_l) entropies for expressions (E) and for characters (Ch) are shown.

	C_1	C_2	A_l(E)	A_l(Ch)
DB	2.81	2.81	7.13	3.05
F0 means, variance	1.57	**1.43**	3.27	1.94
Pitch	**1.56**	1.56	3.24	1.97
Power	2.31	2.36	3.52	2.25
Pitch - Power	2.29	2.35	3.74	2.29
JShimm	2.55	2.48	3.41	2.18
Rhythm	2.12	2.20	3.20	1.89
Rhythm - Pitch	1.70	1.60	3.12	1.78
Rhythm - Pitch - JShimm	1.67	1.64	3.10	1.75
MFCCiVec	2.67	2.65	3.16	1.80
F0iVec	2.12	2.06	3.48	2.08
PoweriVec	2.64	2.65	3.53	2.15
sylDuriVec	2.24	2.81	4.80	2.28
iVecC	2.63	2.59	3.13	1.92
Rhythm - iVecC	2.16	2.29	**3.04**	**1.72**
Rhythm - JShimm - iVecC	2.37	2.18	3.09	1.81

Table 1 shows the results for the objective cluster evaluation. There is a lot of difference between the features and between the corpora. The bold marked values are the best results obtained. For the laboratory recorded corpora the best results are obtained using just F0 or the Pitch combination. While the best results for the audiobook were obtained using the Rhythm and the i-vector combination. Rhythm seems to be more important for the audiobook than for the laboratory corpora. In fact, the Rhythm and Pitch and the Rhythm, Pitch and JShimm combinations achieve almost the same results as the Rhythm and i-vector combination. On the other hand, Rhythm alone performs worse than i-vectors alone, although better than Pitch and JShimm alone.

On the other side, i-vectors do not perform well applied to the laboratory corpora. It seems to be due to the fact, that the laboratory corpora have only one speaker each, while the audiobook has many speakers (although only approximated by imitation). The best results here were obtained using the Pitch parameters. Also on the i-vectors side the best results were obtained with the F0 based i-vectors.

An interesting observation can be made examining closely the individual cluster results for the female laboratory speaker using the i-vectors based on syllable durations. Several clusters of approximate size of 30 to 40 utterances appear to be totally homogeneous, i.e. all labels in these clusters belong to the same emotion ($entropy = 0$), e.g. *angry*, *surprise disgust*, etc. This distribution suggests that the female laboratory speaker uses rhythm as an important tool to communicate emotions. However, this is not true for the male laboratory speaker nor for the audiobook reader.

On the other hand, the clusters are often formed of emotions which acoustically could belong together, such as *joy*, *angry* and *surprise*, or *sad* and *fear*. Although some emotions, specially *fear* and *surprise* were often co-appeared with other emotions. This is not surprising since these emotions can easily

combine with others, for instance one can be surprised positively, i.e. joyful, or negatively, with fear or anger. Also fear can be more aggressive, i.e. angry, or more neutral, or close to sadness.

6. Perceptual Results

Table 2 shows the results for the perceptual experiment. Each number represents the percentage of how often a voice is chosen to represent a book character. Bold numbers indicate the highest preferences. A total of 11 subjects have participated in the experiment, 8 of them not familiar with speech technology.

Table 2: Relative preferences for the voices v0-v9 over the whole paragraph for the narrator and the two present characters.

	v0	v1	v2	v3	v4
$Narrator$	**0.42**	0.06	0.00	0.03	0.04
$Ch2$	0.13	**0.16**	0.14	**0.23**	0.03
$Ch3$	**0.18**	0.13	0.13	**0.31**	0.00
	v5	v6	v7	v8	v9
$Narrator$	**0.23**	0.04	0.10	0.06	0.01
$Ch2$	0.09	0.05	0.03	0.10	0.03
$Ch3$	**0.18**	0.00	0.00	0.00	0.05

The participants had no examples of how the real audiobook characters sounded, so it was their choice of how to interpret the characters. This surely is influenced by the fact whether the participant did or did not know the book, and also by her imagination of how a certain character should sound. Nevertheless, certain voices are systematically preferred for certain characters, and also for different parts of the dialogue. So for instance, the narrator voice is chosen differently for the beginning of the dialogue and for the middle part, where tension rises. The characters are being interpreted more freely, specially the second one. Although more than the half of all voices are chosen randomly, it does not mean that some can not represent the characters adequately. In general, the first 4 (V0-V3) and the 6th voice (V5) were mostly preferred for the interpretation, the first 4 were chosen by the distance calculation. None of the participants selected the neutral voice for all sentences (V4), although it had higher segmental quality. This suggests that the clustering was successful.

7. Conclusions

This work has studied the usefulness if i-vectors in comparison to other features for expressive or emotional speech analysis. For this purpose two different databases were used, one recorded in a laboratory, aimed to imitate certain emotions, and the other one being an audiobook recorded by a professional reader. For each utterance of each database a set of features were extracted, more traditional, prosody based features on the one hand, and i-vector based features, prosodic and spectral, on the other hand. Then, k-means clustering was applied to the corpora using different feature combinations, and the homogeneity of each cluster was evaluated by means of entropy. Additionally, a perceptual experiment was conducted, where the participants had to edit a small paragraph from the audiobook using synthetic voices trained from clusters created with the best feature set for the audiobook.

From the objective results several things can be concluded about the suitability of different features for the expressive speech. First, for the speech corpora recorded in the controlled laboratory environment it seems that the more traditional features, specially the pitch related features, worked best. This is not true though for the audiobook, which was also recorded in studio environment, however, the interpretation of characters and emotions was the choice of the reader. In this case i-vectors seem to be very useful, probably as a consequence of the presence of different speakers (book characters), though only imitated. Probably, in a real multi-speaker database, where the speakers are not imitated, the i-vectors could be even more effective.

Also, even in a controlled environment, the speakers use different means to transmit emotions, as can be seen in the case of the syllable duration based i-vectors. The female speaker obviously used different rhythms for different emotions, while the male speaker did not. As the results suggest, the audiobook reader also made use of different rhythmic patterns for his interpretation, although in a different manner then the female speaker.

The results achieved in the perceptual experiment show that the clustering on the proposed feature set worked well enough as to create believable voices and expressions for given audiobook characters. On the other hand, the interface used for the experiment gives the participants the possibility to be creatively productive and create or edit audiobooks.

The present work shows that, among the studied features, there is no universal one that can be used for all expressive speech analysis. There are important differences between databases, speaker and expressions and speaking styles. Also, the prosodic and acoustic tools used to express emotions can very a lot depending on the speaker even for the same type of expressiveness and context.

8. Acknowledgements

This work was supported by the Spanish Ministerio de Economía y Competitividad and European Regional Development 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.

9. References

[1] 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.

[2] 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.

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

[4] 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.

[5] E. Szekely, J. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Proceedings of Interspeech*, 2011, pp. 2409–2412.

[6] 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.

[7] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. Gales, and K. Knill, "Unsupervised clustering of emotion and

voice styles for expressive TTS," in *Proceedings of ICASSP*, 2012, pp. 4009–4012.

[8] R. Barra-Chicote, J. Yamagishi, S. King, J. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, pp. 394–404, 2010.

[9] 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

[10] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proceedings of ICSLP*, 1996, pp. 1137–1140.

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

[12] 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.

[13] 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.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.4.07)," 2015, http://www.praat.org/.

[15] R. Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, no. 2, pp. 4–29, 1984.

[16] "Speech signal processing toolkit ver. 3.6."

[17] C. Shannon, "A mathematical theory of communication." *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[18] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," *Machine Learning*, vol. 55, no. 3, pp. 311–331, 2004.

[19] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: Design and collection of a multilingual emotional speech database." in *LREC*, 2002.