Proceedings of the 10th International Conference on
Advances in Quantitative Laryngology, Voice and Speech Research
June 3-4, 2013, Cincinnati, Ohio, USA

UNSUPERVISED CLUSTERING OF EXTREME VOCAL EFFECTS

Oriol Nieto
Music and Audio Research Lab, New York University, New York, NY, USA

***Keywords*: Singing Voice; Extreme Vocal Effects; Machine Learning; Clustering**

INTRODUCTION

It is common for singers in music genres such as hard rock or metal to use special voice techniques colloquially known as screams or growls to enhance their expressivity and add a singularly distinctive color to their songs. These techniques are produced by incorporating non-linearities to the vocal folds and vocal tract, which yield harmonically richer (and sometimes noisier) spectrums. Following the work in [1], we call these techniques Extreme Vocal Effects (EVEs), and in this paper we combine a number of acoustical features to automatically classify different singing voice signals into a predefined set of EVEs following a fast and entirely unsupervised process. The automatic classification of these effects could be useful to identify different types of EVEs in, e.g. karaoke systems or educational games[1], where a singer has to produce a specific type of EVE and she or he is evaluated in real-time.

The study of EVEs has been discussed in the digital audio effects literature, where an algorithm for rough and growl can be found in [2]. A study on the production and synthesis of different EVEs can be found in [1]. Under the medical and physiological frameworks, these EVEs have also been considered, with two good examples in [3] and [4].

Given the acoustic similarity of some of these EVEs with various voice disorders, we hypothesize that these features could also be used to classify recordings of patient voices into different disorder categories, following works such as [5].

TYPES OF EXTREME VOCAL EFFECTS

In this work we aim to classify singing voice signals into four different categories: three EVEs (growl, fry scream, and roughness), and the normal singing voice. In this section we describe these three types of EVEs[2].

Growl

This type of EVE is common in death metal and other extreme music genres. It is particularly noisy and the fundamental frequency is rarely perceived, since the sound from the vocal folds is distorted by thorough movements of supra glottal tissues. Growls are usually loud and produce a high amount of spectral variation.

Fry Scream

This scream is similar to the growl, but its spectrum is *brighter* and it is not usually as loud. The fry scream is produced by a series of spaced glottal pulses that can be induced by either exhaling or inhaling. When inhaling, this EVE is popularly known as the "pig squeal".

Roughness

By adding smaller variations on the vocal tract, we can obtain a harmonically richer spectrum while keeping the fundamental frequency perceivable. This type of EVE is more common in hard rock than in extreme metal music.

METHOD

In order to cluster the previously described EVEs we have to extract a specific set of features from the audio signal that capture the differences between the EVEs. We obtain a set of observations by extracting these features from a collection of audio samples. These observations will be the input of an unsupervised machine learning algorithm (*k*-means) that will learn a dictionary of centroids to cluster new observations. In this section we report the details of this process.

Feature Extraction

The features used in this project are: Mel Frequency Cepstral Coefficients (MFCCs), spectral contrast, number of zero crossings in the time domain and loudness. Before extracting these features, we downsample the audio signal to 11025Hz and make it mono. Then, we compute the spectrogram using a Blackman window of 400ms and a hop size of 0.14ms. For each time frame of 400ms, we extract a set of features that will become an observation and will be classified as a specific voice type.

MFCCs have been widely used for both speech and music applications [6], and they mostly capture timbral properties of the audio. After removing the first one, we use the following 13 coefficients in this project, as it is common in the literature [7]. The spectral contrast captures the amount of contrast that exists across a set of octaves in a given spectrum, and it has proven to be a good feature to classify music [8]. In our project we check for 6 octaves, starting from an A1 pitch (55Hz). The number of zero crossings in the time domain is a reliable

[1] Specific example: http://screaminator.urinieto.com/

[2] Audio examples:
https://files.nyu.edu/onc202/public/aql2013/

feature to discriminate between speech and music [9], and it is simply obtained by counting the number of times that the audio signal crosses the zero value. Finally, the loudness is computed using the Root-Mean-Squared (RMS) value of the frequency spectrum.

Learning the Dictionary

To learn a dictionary in an unsupervised process we need enough observations for the clustering algorithm to find centroids that capture differences in the audio signal as generically as possible. Each observation is a feature vector of size *M* representing the set of features for a given time frame. In our case, *M=16* since we have 13 MFCCs, spectral centroid, number of zero crossings and RMS. We use a training dataset containing approximately 150 observations for each type of effect, and define the input matrix of training data as a concatenation of all the time frames of all the audio files from the training dataset (see Fig. 1). The input matrix is standardized to unit variance for each feature. Then we apply *k*-means using the Euclidean distance to learn the dictionary *D*, which will contain *K* different centroids representing the types of singing voice we want to cluster (*K=4:* normal, growl, fry scream and roughness). Consequently, *D* is a *K* x *M* matrix. Notice that the data of the training set is not labeled, making this a completely unsupervised process.

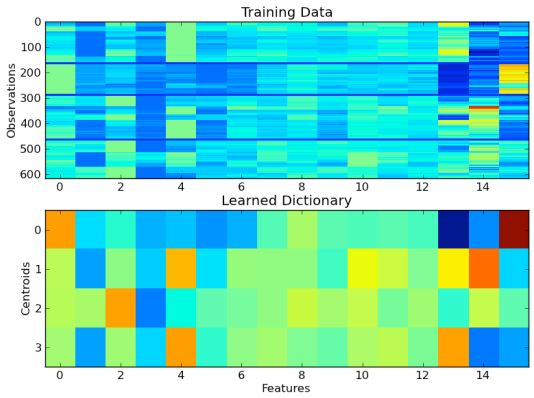

Fig. 1: Training data on top. The horizontal lines are manually added to visualize the differences between clusters. Output dictionary with *K*=4 on the bottom. The first 13 features are MFCCs, then spectral centroid, number of zero crossings, and RMS.

Clustering New Observations

We use the dictionary *D* to classify new observations by finding the centroid that has the minimum Euclidean distance to the features of the new observation. This centroid will be a number between 0 and 3 that will represent the type of vocal effect that this observation is more likely to belong to. This process is fast enough to be able to run in a real time application.

Results And Discussion

This process has been tested on a dataset containing 246 labeled observations, with approximately 60 observations for each vocal effect. The amount of correctly clustered observations is 92.7%. Both the training and testing datasets were sung and collected by the author, so the learned dictionary is highly overfitted to the author's voice. In order to obtain a more generic dictionary we would need audio samples from a many great number of singers. Since this is an unsupervised process, the more data in the dataset the better results we should obtain.

Conclusion

We have discussed a set of features that, together, can classify audio recordings into a set of vocal effects in a real-time, unsupervised process. As future work, we could use these features to help identify EVEs in music tracks, similarly to [10]. Finally, we believe this process could help classify patient recordings into different voice disorders, since EVEs are acoustically similar to some of these disorders.

Acknowledgments

Thanks to "Caja Madrid" foundation for the funding and Eric J. Humphrey for the good advice.

References

[1] Nieto, O. (2008). *Voice Transformations for Extreme Vocal Effects*. Master's Thesis. Pompeu Fabra University.

[2] Loscos, A., & Bonada, J. (2004). *Emulating Rough And Growl Voice In Spectral Domain*. Proc. of the 7th International Conference on Digital Audio Effects. Naples, Italy.

[3] McGlashan, J., Sadolin, C., & Kjelin, H. (2007). *Can Vocal Effects Such As Distortion, Growling, Rattle and Grunting Be Produced Without Traumatising The Vocal Folds?* Proc. of the Pan European Voice Conference. Groningen, The Netherlands.

[4] Eckers, C., Hütz, D., Kob, M., Murphy, P., Houben, D., & Lehnert, B. (2009). *Voice Production in Death Metal Singers*. In NAG/DAGA International Conference on Acoustics (pp. 1747–1750). Rotterdam, The Netherlands.

[5] Childers, D. G. (1990). *Speech Processing And Synthesis For Assessing Vocal Disorders*. IEEE Engineering in Bedicine and biology Magazine, 9(1), 69–71.

[6] Logan, B., & Salomon, A. (2001). *A Music Similarity Function Based on Signal Analysis*. Proc. of the IEEE International Conference on Multimedia and Expo. Tokyo, Japan.

[7] Tzanetakis, G., & Cook, P. (2002). *Musical Genre Classification of Audio Signals*. IEEE Transactions on Speech and Audio Processing, 10(5), 293–302.

[8] Lu, L., Zhang, H., Tao, J., & Cui, L. (2002). *Music Type Classification by Spectral Contrast Feature*. Proc. of the IEEE International Conference on Multimedia Expo (pp. 113–116). Lausanne, Switzerland.

[9] Saunders, J. (1996). *Real-time Discrimination of Broadcast Speech Music*. Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (pp. 993–996). Atlanta, GA, USA.

[10] Regnier, L., & Peeters, G. (2009). *Singing Voice Detection In Music Tracks Using Direct Voice Vibrato Detection*. Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (pp. 1685–1688). Taipei, Taiwan.