Single-speaker/multi-speaker co-channel speech classification

Stéphane Rossignol[1], *Olivier Pietquin*[1]

[1]IMS research group, SUPÉLEC – Metz Campus, F-57070 Metz

stephane.rossignol@supelec.fr, olivier.pietquin@supelec.fr

Abstract

The demand for content-based management and real-time manipulation of audio data is constantly increasing. This paper presents a method to identify temporal regions, in a segment of co-channel speech, as being either single-speaker or multi-speaker speech. The state of the art approach for this purpose is the kurtosis. In this paper, a set of complementary time-domain and frequency-domain features is studied. The employed classification scheme is the one-class SVM classifier. A recognition rate of 94.75 % is reached. The set of features providing the best performance is determined.

Index Terms: Speech segmentation, Speaker characterization and recognition

1. Introduction

When more than one speaker are active simultaneously, speech analysis tools, such as speech recognition systems, speaker diarization systems (see [7]), etc., become generally unreliable. The identification of two speech classes, that is to say single-speaker speech signals and multi-speaker speech signals, is thus a compulsory preliminary step for these applications. We propose a new approach to distinguish single-speaker speech from multi-speaker speech. The approach is composed of two steps. First, a set of features that measure distinct properties of single-speaker speech signals and multi-speaker speech signals is extracted. Here, our focus is on the extraction of meaningful features. Second, these features are combined into a multidimensional classification framework, which is the one-class SVM classifier. Two similarity measures are tested. Various combinations of features are studied. Results on real data are given.

2. Related works

Local kurtosis measures are usually used to perform the classification of single-speaker/multi-speaker segments of speech signals. According to the central limit theorem, the distribution of additioned speech signals is supposed to tend towards a Gaussian and thus to have a lower kurtosis than each of the original speech signals ([5], [6]). In this paper, features different from the kurtosis are studied, mainly in order to increase the recognition rate of the single/multi-speaker speech classification.

The features studied in this paper are actually based on work done in the field of the time-domain speech/music segmentation. In this work, the speech class is composed of single-speaker speech and music comprises various musical types, such as monophonic or polyphonic music, classical, rock, etc. In the present paper, the multi-speaker speech class is assumed to “resemble” the music class. In [1] a technique is described which discriminates speech from music. Features based on the zero-crossing-rate (ZCR) and on the energy are used. A trained Multivariate Gaussian classifier is used. Segments are 2.4 seconds long. Experimental results show performance approaching 98 % of correct classification. In [2] a computer system is described which is capable of distinguishing speech signals from music signals. Thirteen features have been evaluated. Several classifiers have been examined. The best classifier classifies with 1.4 % error when 2.4 second segments of sound are considered. In [8] a technique is described for segmenting a sound track into two classes. The features used are cepstral coefficients and their first order derivatives. A Gaussian Mixture Model is used. On a frame-by-frame basis (frames are 26 ms long), the error rate is around 20 %. In [3], on a segment-by-segment basis (segments are 1 second long), an error rate of 1.3 % is obtained. We have decided to use in this paper the features studied in [3]. Similar recognition rates are expected for the segmentation single-speaker/multi-speaker studied in this paper.

Some work has been done as well on multi-speaker voice activity detection. However, in this work, first, more than a single microphone is usually available whereas in the present paper co-channel speech is studied; and, second, it is assumed that one of the speakers predominates over the others, whereas this assumption is not made here. The features used by the existing approaches are therefore either useless for our purpose or they must be modified considerably. Furthermore, this makes performance comparison a difficult task. In [9], the energy is studied; in [10], the cross-channel correlation is used; in [11], the loudness in sub-bands, the energy, the total loudness, the ZCR, etc. are used; in [12], Short Time Fourier Transform (STFT) features are extracted; in [13], Mel-frequency cepstral coefficients (MFCCs), the energy, the ZCR, the kurtosis, the fundamentalness, the cross-channel correlation, etc. are used; in [14], MFCCs, the energy, linear predictive coding (LPC) coefficients, the diarization posterior entropy are studied. Recognition rates higher than 80 %/90 % are commonly reported.

3. The algorithm

3.1. First step: feature extraction

In this paper, nine features are examined. They are based on the following short-term characteristics ([1] and [2]):

- **The spectral flux**, which is defined as the $L2$-norm of the difference between the magnitude of the STFT spectrum evaluated at two successive sound frames. Notice that each evaluation of the STFT is normalized in energy.
- **The spectral centroid** f_c, which is defined as the “balancing point” of the magnitude of the STFT spectrum $\hat{S}$ evaluated at a sound frame:
$$f_c = \frac{\sum_{i=0}^{N/2} |\hat{S}(i)| f_i}{\sum_{i=0}^{N/2} |\hat{S}(i)|}$$
where f_i is the frequency at the ith bin and where N is the size of the FFT (Fast Fourier Transform).

- **The zero-crossing rate** (ZCR), which is defined as the number of time-domain zero-crossings within a frame.
- **The cepstrum resynthesis residual magnitude** (CRRM), which is defined as the $L2$-norm of the difference between the magnitude of the STFT spectrum $\hat{S}$ evaluated at a sound frame and the "smoothed spectrum" $\hat{M}$ evaluated at the same frame. $\hat{M}$ is obtained using the real cepstrum $\hat{C}$:
$$\hat{C} = \text{real}\left(\text{FFT}^{-1}\left(\log\left(|\hat{S}|\right)\right)\right)$$
$$\hat{C}' = \hat{C}W$$
$$\hat{M} = \exp\left(\text{FFT}\left(\hat{C}'\right)\right)$$
where W is a window with value 1 or 0. Only the n first coefficients (corresponding to smallest positive frequencies) and the n latest coefficients (corresponding to their negative counterpart) of $\hat{C}$ are kept. A 2048-point FFT is performed. For a 44.1 kHz sampling rate, $n = 50$ is used. $\hat{M}$ is essentially a low pass filtering of the spectrum $\hat{S}$, so that a better fit for noise signals than for harmonic signals is obtained.
- **The kurtosis**, which is defined as $\kappa_x = \text{E}\left[x^4\right]/\text{E}^2\left[x^2\right]$, where $x = s - \text{E}[s]$ and s corresponds to the samples on a sound frame.

Frames are 20 ms long and are overlapped by 50 %. The nine features extracted from these short-term characteristics are:

- **features 1 & 2:** mean and logarithm of the variance of the spectral flux
- **features 3 & 4:** mean and logarithm of the variance of the spectral centroid
- **features 5 & 6:** mean and logarithm of the variance of the ZCR
- **features 7 & 8:** mean and logarithm of the variance of the CRRM
- **feature 9:** mean of the kurtosis

Means and variances are computed in a one-second segment. Segments are overlapped by 50 %.

3.2. Second step: classification

3.2.1. SVM

In order to identify the segments, one-class SVM classifiers are used. More details about the SVM classifier can be found in [15]. For computational efficiency, one-class SVMs are preferred to the classical two-class SVM. Indeed, when the training set comprises a large number of samples, the computation time required to train the two-class SVM classifier becomes unbearable.

In this paper, only the RBF gaussian kernel k has been used.

A one class SVM is trained for each of the two classes and, for a segment to classify, a similarity measure is used to indicate in which class the segment is likely to belong.

3.2.2. Similarity measures

Consider x_k a sample to be classified; $x_i^{1\cdot}$ the $m^{1\cdot}$ training samples, and $\alpha_i^{1\cdot}$ and $b^{1\cdot}$ the trained parameters, for the first class; $x_i^{2\cdot}$ the $m^{2\cdot}$ training samples, and $\alpha_i^{2\cdot}$ and $b^{2\cdot}$ the trained parameters, for the second class.

The most obvious similarity measure is to compare $y_k^{1\cdot}$ and $y_k^{2\cdot}$, where:

$$y_k^{1\cdot} = \sum_{i=1}^{m^{1\cdot}} \alpha_i^{1\cdot} k\left(x_i^{1\cdot},\ x_k\right) + b^{1\cdot}$$

$$y_k^{2\cdot} = \sum_{i=1}^{m^{2\cdot}} \alpha_i^{2\cdot} k\left(x_i^{2\cdot},\ x_k\right) + b^{2\cdot}$$

If $y_k^{1\cdot} > y_k^{2\cdot}$, it is decided that x_k belongs to the first class and conversely.

However, it has been shown that more efficient similarity measures exist. Such a dissimilarity measure is described in detail in [16] and [17]. We tested it in this paper. For the first class, the dissimilarity is equal to:

$$I_k^{1\cdot} = \frac{|\widehat{c^{1\cdot}c_k}|}{|\widehat{c^{1\cdot}p^{1\cdot}}|}$$

where:

$$\widehat{c^{1\cdot}c_k} =$$

$$\arccos\left(\frac{\sum_{i=1}^{m^{1\cdot}} \alpha_i^{1\cdot} k\left(x_i^{1\cdot}, x_k\right)}{\left(\sum_{i=1}^{m^{1\cdot}}\sum_{j=1}^{m^{1\cdot}} \alpha_i^{1\cdot}\alpha_j^{1\cdot} k\left(x_i^{1\cdot}, x_j^{1\cdot}\right)\right)^{0.5} \left(k(x_k, x_k)\right)^{0.5}}\right)$$

and:

$$\widehat{c^{1\cdot}p^{1\cdot}} = \arccos\left(\frac{-b^{1\cdot}}{\left(\sum_{i=1}^{m^{1\cdot}}\sum_{j=1}^{m^{1\cdot}} \alpha_i^{1\cdot}\alpha_j^{1\cdot} k\left(x_i^{1\cdot}, x_j^{1\cdot}\right)\right)^{0.5}}\right)$$

If $I_k^{1\cdot} < I_k^{2\cdot}$, it is decided that x_k belongs to the first class and conversely.

4. Experimentation

4.1. Data

To assess the performance of the above-mentioned features, the standard BREF80 database has been used: see [4]. This database contains several hours of speech uttered by 90 different native French speakers (50 females and 40 males). In order to obtain multi-speaker speech signals, the BREF80 signals are added each other. The number of added signals varies between two and ten. The signals to mix are randomly chosen from BREF80, the amplitude of each of them being uniformly chosen between 0.3 and 1 before adding. The features mentioned above have been computed both on the original BREF80 signals (single-speaker speech) and on the mixed signals (multi-speaker co-channel speech).

Two data sets are considered. Each of them is composed of 26850 segments (12316 single-speaker segments; 14534 multi-speaker segments). The first set is used to train the classifiers. The second one is used to cross-test the training stage. The mean of each feature of the first set is normalized to 0 and its standard deviation to 1. The second set is normalized using the normalization coefficients obtained for the first set.

4.2. Motivations

Single-speaker speech is a succession of noise periods, such as unvoiced consonants, and of periods of relative stability, like vowels. Multi-speaker speech signals can be rather regarded as a rapid succession of mixtures of voiced and unvoiced components. Therefore, the selected features give very different values for voiced and unvoiced single-speaker speech; and they are relatively constant within a segment of multi-speaker speech. The variances should be higher for single-speaker speech than for multi-speaker speech. Due to the high number of transitions in the multi-speaker speech signals, the spectral flux mean should tend to be greater for these signals than for single-speaker speech. The more mixed sources there are, the higher is the probability that, at a given time, a voiced component is present. The mean of the CRRM should thus rapidly increase with the number of mixed sources, as voiced signals are better fitted than noise signals. Similarly, the mean of the centroid should rapidly decrease with the number of mixed sources, as the energy of the voiced components is gathered in the lowest frequencies; and the mean of the ZCR should rapidly decrease with the number of mixed sources.

4.3. Behaviour of each feature

In figures 1, 2, 3, 4, and 5, the behaviour of the studied features is examined.

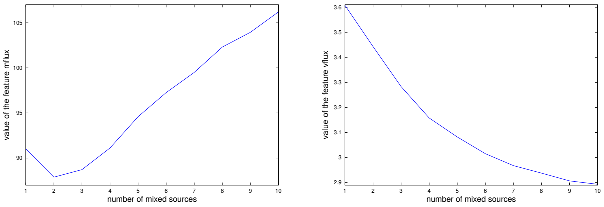

Figure 1: Behaviour of the mean (left) and of the variance (right) of the spectral flux; x-axis: number of mixed sources; y-axis: value of the feature.

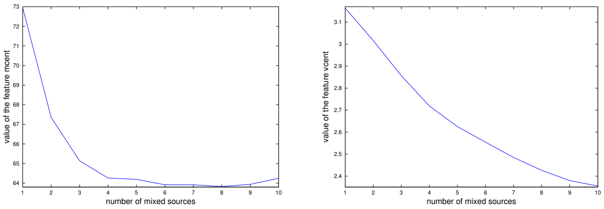

Figure 2: Behaviour of the mean (left) and of the variance (right) of the centroid; x-axis: number of mixed sources; y-axis: value of the feature.

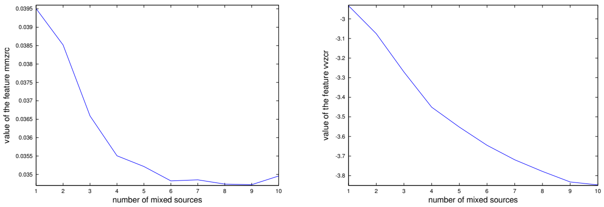

Figure 3: Behaviour of the mean (left) and of the variance (right) of the ZCR; x-axis: number of mixed sources; y-axis: value of the feature.

It can be noticed that the features behave as expected and that each of them can be used to perform efficiently the segmentation single-speaker/multi-speaker speech. However, the mean of the spectral flux is lower when 2 or 3 sources are present than for single-speaker signals. This is due to the fact that the increase in the amount of transitions is counterbalanced by the increase in the amount of frames which are voiced rather than unvoiced. When the number of mixed sources continues to increase, the effect of the transitions (that is to say, higher values for the flux) becomes prominent. Furthermore, the kurtosis does not seem to perform better than the other features.

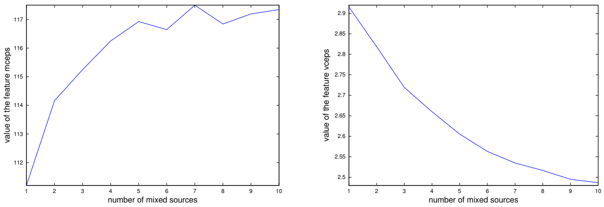

Figure 4: Behaviour of the mean (left) and of the variance (right) of the CRRM; x-axis: number of mixed sources; y-axis: value of the feature.

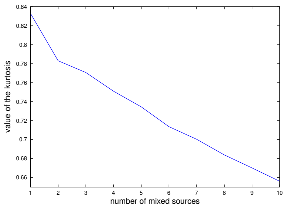

Figure 5: Behaviour of the mean of the kurtosis; x-axis: number of mixed sources; y-axis: value of the feature.

4.4. Performance of the classification

The percentage of correctly classified segments is given in tables 1, 2 and 3. Table 1 shows performance obtained when using the features individually. Table 2 presents performance obtained when using each pair of features. Table 3 shows performance obtained when using various feature sets. In each table, columns 2 and 3 list training errors obtained by comparing the results of the segmentation to be found and the results of the training step, using respectively the first similarity measure and the second one. For the cross-testing, on columns 4 and 5, the classifiers are trained using the set number 1, and the test is operated on the set number 2, respectively using the first similarity measure and the second one. The SVM parameters, that is to say here the "σ^2" of the RBF gaussian kernel, need to be carefully fine-tuned. This process, which is quite time consuming, has been performed for the feature sets considered in the three tables. The parameter ν has been fixed to 0.1, which is a commonly used value.

	training		cross-testing	
measure:	*1st*	*2nd*	*1st*	*2nd*
feature 1	63.49 %	63.49 %	63.01 %	63.49 %
feature 2	85.25 %	85.29 %	**85.55 %**	**85.5 %**
feature 3	67.4 %	67.7 %	67.68 %	68.7 %
feature 4	78.57 %	78.57 %	78.01 %	77.99 %
feature 5	68.09 %	68.89 %	67.47 %	68.65 %
feature 6	74.03 %	74.05 %	73.28 %	73.21 %
feature 7	65.78 %	66.84 %	66.62 %	67.45 %
feature 8	77.96 %	78.09 %	79 %	79.02 %
kurtosis	71.85 %	71.7 %	71.57 %	72.56 %

Table 1: Percentage of correctly classified segments when using each feature individually

Considering the three tables, the differences in terms of recognition rate between the training and the cross-testing are small, indicating that the classification scheme in use as a good generalization behaviour.

Table 1 shows that the retained features are useful to separate the two classes. The best feature is the logarithm of the variance of the spectral flux. The kurtosis performance is in the middle range. The worst feature is the mean of the spectral flux: this is consistent with the comment provided in section 4.3

concerning this feature). The second similarity measure is often more efficient than the first one.

	training		cross-testing	
measure:	*1st*	*2nd*	*1st*	*2nd*
flux	87.45 %	87.25 %	**87.51 %**	**87.12 %**
centroid	80.64 %	80.9 %	80.62 %	80.43 %
ZCR	85.51 %	85.55 %	85.2 %	85.33 %
CRRM	82.20 %	82.18 %	83.80 %	83.76 %

Table 2: Percentage of correctly classified segments for each pair of features, that is to say the mean and log(variance) of each short-term characteristic

Tables 2 and 3 indicate that when more than a single feature is used, the two similarity measures provide similar performance. The flux provides the best pair of features. Variances give slightly better results than means. The best recognition rate is obtained using the 9 features together: a recognition rate of 94.21 % for the cross-testing is obtained.

Considering both similarity measures, the best feature sets and the corresponding recognition rates (cross-testing only) are given in table 4. For both similarity measures, the kurtosis does not appear in the set of features providing the best performance. Most of the features are means. It must be noticed that, for computational time reasons, the "σ^2" parameters could not be optimized for each of the 511 possible feature sets. However, using reasonable values for the "σ^2"s, a recognition rate of 94.75 % is reached, improving table 3 best performance (94.21 %).

	training		cross-testing	
measure:	*1st*	*2nd*	*1st*	*2nd*
4 variances	87.54 %	87.36 %	88.31 %	88.34 %
4 means	87.02 %	87.00 %	86.61 %	86.69 %
kurtosis + 4 variances	87.94 %	87.95 %	88.36 %	88.59 %
kurtosis + 4 means	87.88 %	87.82 %	87.86 %	87.99 %
9 features	94.19 %	95.16 %	93.80 %	**94.21** %

Table 3: Percentage of correctly classified segments when using various sets of features

	features	recognition rate
SVM – 1st measure	1 2 3 4 5 6 7	94.28 %
SVM – 2nd measure	1 2 3 5 6 7	**94.75** %

Table 4: Best feature sets for both similarity measures and corresponding recognition rates

5. Conclusion

A system identifying temporal regions in a segment of co-channel speech, as being either single-speaker or multi-speaker speech, is presented in this paper. Nine features have been tested. Each of them has been validated (recognition rates between 63.49 % and 85.5 %). Furthermore, the performance of the system when using various feature sets have been examined, showing that using combination of features allows to improve the performance. The best recognition rate obtained is 94.75 %. It has been obtained using the feature set [1 2 3 5 6 7].

The one-class SVM classifier provides a parsimonious parametrization of the training set, as most of the αs are equal to 0. As our goal is to build real-time tools for audio segmentation and manipulation, this is of interest for us.

Some improvements and future work are envisioned. For instance, some features that have been mentioned in the second section have not been evaluated here. Furthermore, the optimization of the SVM parameters has not been fully achieved.

This work, which is a first step for the diarization of speech signals into single-speaker/multi-speaker classes, provides very promising results.

6. References

[1] Saunders J. "Real-time discrimination of broadcast speech/music", IEEE Trans. on Acoustics, Speech, and Signal Processing, 993-995, 1996.

[2] Scheirer E. and Slaney M. "Construction and evaluation of a robust multifeatures speech/music discriminator", IEEE Trans. on Acoustics, Speech, and Signal Processing, 1331-1334, 1997

[3] Rossignol S., Rodet X., Soumagne J., Collette J.-L. and Depalle P. "Automatic characterisation of musical signals: feature extraction and temporal segmentation", Journal of New Music Research, 28:4, 281-295, 1999

[4] Lamel L. F., Gauvain J.-L. and Eskenazi M. "BREF, a Large Vocabulary Spoken Corpus for French", Eurospeech, 1991

[5] LeBlanc J.P and De Leòn P. L. "Source Separation of Speech Signals using Kurtosis Maximization", IEEE Trans. on Acoustics, Speech, and Signal Processing, 1029-1032, 1998

[6] Krishnamachari K. R., Yantorno R. E. and Lovekin J. M. "Use of local kurtosis measure for spotting usable speech segments in co-channel speech", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2001

[7] Gutzwiller J.-L., Frezza-Buet H. and Pietquin O. "Online Speaker Diarization with a Size-Monitored Growing Neural Gas Algorithm", Proc. 18th European Symposium on Artificial Neural Networks, 6 pages, 2010

[8] Seck M., Bimbot F., Zugaj D. and Delyon B. "Two-class signal segmentation for speech/music detection in audio tracks", Eurospeech, 1999

[9] Bertrand A. and Moonen M. "Energy-based multi-speaker voice activity detection with an ad-hoc microphone array", Proc. IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP), 2010

[10] Laskowski K., Jin Q. and Schultz T. "Crosscorrelation-based Multispeaker Speech Activity Detection", Proc. International Conference of Spoken Language Processing (ICSLP), 2004

[11] Pfau T., Ellis D. P. W. and Stolcke A. "Multispeaker speech activity detection for the ICSI meeting recorder", Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 2001

[12] Checka N., Wilson K. W., Siracusa M. R. and Darrell T. "Multiple person and speaker activity tracking with a particle filter", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2004

[13] Wrigley S. N., Brown G. J., Wan V. and Renals S., "Feature selection for the classification of crosstalk in multi-channel audio" Eurospeech, 2003

[14] Boakye K., Trueba-Hornero1 B., Vinyals O. and Friedland G. "Overlapped speech detection for improved speaker diarization in multiparty meetings", Proc. IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP), 2008

[15] Schölkopf B. and Smola A. J. "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond", the MIT Press, Cambridge, MA, 2002

[16] Desobry F., Davy M. and Doncarli C. "An Online Kernel Change Detection Algorithm", IEEE Trans. on Signal Processing, 2005

[17] Rossignol S. and Davy M. "Détection de ruptures à l'aide des SVM 1 classe pour la segmentation des signaux sonores musicaux", GRETSI, 2007