Hypothesis Testing as a Performance Evaluation Method for Multimodal Speaker Detection

Patricia Besson[1] and Murat Kunt[1]

Signal Processing Institute (ITS), Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland

Abstract. This work addresses the problem of detecting the speaker on audio-visual sequences by evaluating the synchrony between the audio and video signals. Prior to the classification, an information theoretic framework is applied to extract optimized audio features using video information. The classification step is then defined through a hypothesis testing framework so as to get confidence levels associated to the classifier outputs. Such an approach allows to evaluate the whole classification process efficiency, and in particular, to evaluate the advantage of performing or not the feature extraction. As a result, it is shown that introducing a feature extraction step prior to the classification increases the ability of the classifier to produce good relative instance scores.

1 Introduction

This work addresses the problem of detecting the current speaker among two candidates in an audio-video sequence, using a single camera and microphone. To this end, the detection process has to consider both the audio and video clues as well as their interrelationship to come up with a decision. In particular, previous works in the domain have shown that the evaluation of the synchrony between the two modalities, interpreted as the degree of mutual information between the signals, allowed to recover the common source of the two signals, that is, the speaker [1], [2].

Other works, such as [3] and [4], have pointed out that fusing the information contained in each modality at the feature level can greatly help the classification task: the richer and the more representative the features, the more efficient the classifier. Using an information theoretic framework based on [3] and [4], audio features specific to speech are extracted using the information content of both the audio and video signals as a preliminary step for the classification. Such an approach and its advantages have already been described in details in [5]. This feature extraction step is followed by a classification step, where a label ”speaker” or ”non-speaker” is assigned to pairs of audio and video features. The definition of this classification step constitutes the contribution of this work.

As stated previously, the classifier decision should rely on an evaluation of the synchrony between pairs of audio and video features. In [4], the authors formulate the

* This work is supported by the SNSF through grant no. 2000-06-78-59. The authors would like to thanks Dr. J.-M. Vesin, J. Richiardi and U. Hoffmann for fruitful discussions.

Besson P. and Kunt M. (2006).
Hypothesis Testing as a Performance Evaluation Method for Multimodal Speaker Detection.
In *Proceedings of the 2nd International Workshop on Biosignal Processing and Classification*, pages 106-115
DOI: 10.5220/0001224701060115

evaluation of such a synchrony as a binary hypothesis test asking about the dependence or independence between the two modalities. Thus, a link can be found with mutual information which is nothing else than a metric evaluating the degree of dependence between two random variables [6]. The classifier in [4] ultimately consists in evaluating the difference of mutual information between the audio signal and video features extracted from two potential regions of the image. The sign of the difference indicates the video speech source. We have taken a similar approach in [5], showing that such a classifier fed with the previously optimized audio features leads to good results.

In the present work, the classification task is cast in a hypothesis testing framework as well. The objective however is to define not only a classifier, but the means for evaluating the multimodal classification chain performance. To this end, the hypothesis tests are defined using the Neyman-Pearson frequentist approach [7] and one test is associated to each potential mouth region. This way, the ability of the classifier to produce good relative instance scores can be measured. Moreover, an evaluation of the whole classification process, including the feature extraction step, can be introduced. It allows to assess the benefit of optimizing features prior to performing the classification.

The paper is organized as follows: sec. 2 introduces the multimodal information theoretic feature extraction framework and explains how it is applied to extract audio features specific to speech. Sec. 3 describes the hypothesis testing approach taken, showing that it comes finally to evaluate the mutual information in each mouth region with respect to a threshold. In the last section, some results are presented. The behavior of the classifier itself is analyzed and a comparative study of the classification chain performance involving optimized and non-optimized audio features is performed.

2 Extraction of Optimized Audio Features for Speaker Detection: Information Theoretic Approach

2.1 Multimodal Feature Extraction Framework

Given different mouth regions extracted from an audio-video sequence and corresponding to different potential speakers, the problem is to assign the current speech audio signal to the mouth region which effectively did produce it. This is therefore a decision, or classification, task.

Let the speaker be modelled as a bimodal source S emitting jointly an audio and a video signal, A and V. The source S itself is not directly accessible but through these measurements. The classification process has therefore to evaluate whether two audio and video measurements are issued from a common estimated source $\hat{S}$ or not, in order to estimate the class membership of this source. This class membership, modelled by a random variable O defined over the set Ω_O, can be either "speaker" or "non-speaker". Obviously, the overall goal of the classification process is to minimize the classification error probability $P_e = P(\hat{O} \neq O)$, where the wrong class is assigned to the audio-visual features pair. In the present case, a good estimation of the class $\hat{O}$ of the source implies a correct estimation $\hat{S}$ of this source. This source estimate is inferred from the audio and video measurements by evaluating their shared quantity of information. However, these measurements are generally corrupted by noise due

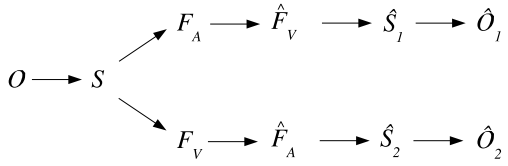


Fig. 1. Graphical representation of the related Markov chains modelling the multimodal classification process.

to independent interfering sources so that the source estimate and thus the classifier performance might be poor.

Preliminarily to the classification, a feature extraction step should be performed in order to possibly retrieve the information present in each modality that originates from the common source S while discarding the noise coming from the interfering sources. Obviously, this objective can only be reached by considering the two modalities together. Now, given that such features F_A and F_V (viewed as random variables hereafter) can be extracted, the resulting multimodal classification process is described by two first order Markov chains, as shown on Fig. 1 [5]. Notice that for the sake of the explanation, the fusion at the decision or classifier level for obtaining a unique estimate $\hat{O}$ of the class is not represented on this graph. F_A and F_V describe specifically the common source and are then related by their joint probability $p(F_A, F_V)$. Thus, an estimate $\hat{F}_V$ of F_V, respectively, $\hat{F}_A$ of F_A, can be inferred from F_A, respectively, F_V. This allows to define the transition probabilities for $F_A \longrightarrow \hat{F}_V$ and $F_V \longrightarrow \hat{F}_A$ (since $p(\hat{F}_V|F_A) = p(\hat{F}_V, F_A)/p(F_A)$, and $p(\hat{F}_A|F_V) = p(\hat{F}_A, F_V)/p(F_V)$). Two classification error probabilities and their associated lower bounds can be defined for these Markov chains, using Fano's inequality [3]:

$$P_{e_1} \geqslant \frac{H(O) - I(F_A, \hat{F}_V) - 1}{\log |\Omega_O|}, \tag{1}$$

$$P_{e_2} \geqslant \frac{H(O) - I(F_V, \hat{F}_A) - 1}{\log |\Omega_O|}, \tag{2}$$

where $|\Omega_O|$ is the cardinality of O, I the mutual information, and H the entropy. Since the probability densities of $\hat{F}_A$ and F_A, respectively $\hat{F}_V$ and F_V, are both estimated from the same data sequence A, respectively V, it is possible to introduce the following approximations: $I(F_A, \hat{F}_V) \approx I(\hat{F}_A, F_V) \approx I(F_A, F_V)$ [3]. Moreover, the symmetry property of mutual information allows to define a joint lower bound on the classification error P_e:

$$P_e = P_{\{e_1, e_2\}} \geqslant \frac{H(O) - I(F_A, F_V) - 1}{\log |\Omega_O|}. \tag{3}$$

$|\Omega_O|$ is supposed to remain fixed during the optimization (only two classes in all cases) and each class is assumed to have the same probability. Consequently, $H(O)$remains constant: $H(O) = -1$. Moreover, $\log |\Omega_O| = 1$, so that Eq. (3) becomes:

$$P_e \geqslant -2 - I(F_A, F_V). \tag{4}$$

To be efficient, the minimization of P_e should therefore include the minimization of the right-hand term of the inequality (4) and therefore, the maximization of the mutual information between the extracted features F_A and F_V corresponding to each modality. However, for the resulting feature sets to compactly describe the relationship between the two modalities, a normalization term involving the joint entropy $H(F_A, F_V)$ has to be introduced, leading to the definition of a feature efficiency coefficient [3]:

$$e(F_A, F_V) = \frac{I(F_A, F_V)}{H(F_A, F_V)} \in [0, 1]. \tag{5}$$

Maximizing $e(F_A, F_V)$ still minimizes the lower bound on the error probability defined in Eq. (4) while constraining inter-feature independencies. In other words, the extracted features F_A and F_V will tend to capture specifically the information related to the common origin of A and V, discarding the unrelated interference information. The interested reader is referred to [3] and [5] for more details.

Applying this framework to extract features, the bound on the classification error probability is minimized. However, there is no guarantee that this bound is reached during the classification process: this depends on the choice of a suitable classifier.

2.2 Signal Representation

Before applying the optimization framework previously described to the problem at hand, both audio and video signals have to be represented in a suitable way.

Physiological evidence points to the motion in the mouth region as a visual clue for speech. The video features are thus the magnitude of the optical flow estimated over T frames in the mouth regions (rectangular regions including the lips and the chin), signed as the vertical velocity component. These mouth regions are roughly extracted using the face detector depicted in [8]. $T-1$ video feature vectors $F_{V,t}$ ($t=1,\ldots,T-1$) are obtained, each element of these vectors being an observation of the random variable F_V.

For the audio representation to describe the salient aspects of the speech signal, while being robust to variations in speaker or acquisition conditions, we use a set of $T-1$ vectors $\boldsymbol{C_t}$, each containing P mel-frequency cepstrum coefficients (MFCCs): $\{C_t(i)\}_{i=1,\ldots,P}$ with $t = 1,\ldots,T-1$ (the first coefficient has been discarded as it pertains to the energy).

2.3 Audio Feature Optimization

The information theoretic feature extraction previously discussed is now used to extract audio features that compactly describe the information common with the video features. For that purpose, the one-dimensional (1D) audio features $F_{A,t}(\boldsymbol{\alpha})$, associated to the random variable F_A are built as the linear combination of the P MFCCs:

$$F_{A,t}(\boldsymbol{\alpha}) = \sum_{i=1}^{P} \boldsymbol{\alpha}(i) \cdot C_t(i) \quad \forall t = 1,\ldots,T-1. \tag{6}$$

Thus, the set of $P \cdot (T-1)$ parameters is reduced to $1 \cdot (T-1)$ values $F_{A,t}(\boldsymbol{\alpha})$. The optimal vector $\boldsymbol{\alpha}$ could be obtained straightaway by minimizing the classification error bound given by Eq. (4). However, a more specific and constraining criterion is introduced here. This criterion consists in the squared difference between the efficiency coefficient computed in two mouth regions (referred to as M_1 and M_2). This way, the discrepancy between the marginal densities of the video features in each region are taken into account. Moreover, only one optimization is performed for two mouths resulting in a single set of optimized audio features. It implies however that the potential number of speakers is limited to two in the test audio-video sequences.

If $F_V^{M_1}$ and $F_V^{M_2}$ denote the random variables associated to regions M_1 and M_2 respectively, then the optimization problem becomes:

$$\boldsymbol{\alpha}_{opt} = \arg\max_{\boldsymbol{\alpha}} \left\{ [e(F_V^{M_1}, F_A(\boldsymbol{\alpha})) - e(F_V^{M_2}, F_A(\boldsymbol{\alpha}))]^2 \right\}. \quad (7)$$

Notice finally that the probability density functions required in the estimation of the mutual information are estimated in a non-parametric way using Parzen windowing.

3 Hypothesis Testing as a Classifier and an Evaluation Tool

3.1 Hypothesis Testing for Classification

The previous section has shown how features specific to the classification problem at hand can be extracted through a multimodal information theoretic framework. The application of this framework results in the minimization of the lower bound on the classification error probability. But the question of reaching the bound itself relies on the choice of a suitable classifier.

Hypothesis tests are used in detection problems in order to take the most appropriate decision given an observation x of a random variable X. In the problem at hand, the decision function has to decide whether two measurements A and V originate from a common bimodal source S - the speaker - or from two independent sources - speech and video noise. As previously stated, the problem of deciding between two mouth regions which one is responsible for the simultaneously recorded speech audio signal can be solved by evaluating the synchrony, or dependence relationship, that exists between this audio signal and each of the two video signals.

From a statistical point of view, the dependence between the audio and the video features corresponding to a given mouth region can be expressed through a hypothesis framework, as follows [4]:

$$H_0 : F_{A,t}, F_{V,t} \sim P_0 = P(F_A) \cdot P(F_V),$$
$$H_1 : F_{A,t}, F_{V,t} \sim P_1 = P(F_A, F_V).$$

H_0 postulates the data to be governed by a probability density function stating the independence of the video and audio sources. The mouth region should therefore be labelled as "non-speaker". Hypothesis H_1 states the dependence between the two modalities: the mouth region is then associated to the measured speech signal and classified as "speaker". The two hypothesis are obviously mutually exclusive.

The Neyman-Pearson approach to hypothesis tests [7] consists in formulating certain probabilities associated with the hypothesis test. The false-alarm probability, or size α of the test, is defined as:

$$\alpha = P(\hat{H} = H_0 | H = H_1), \tag{8}$$

while the detection probability, or power β of the test, is given by:

$$\beta = P(\hat{H} = H_1 | H = H_1). \tag{9}$$

The Neyman-Pearson criterion selects the most powerful test of size α: the decision rule should be constructed so that the probability of detection is maximal while the probability of false-alarm do not exceed a given value α. Using the log-likelihood ratio, the Neyman-Pearson test can be expressed as follows:

$$\Lambda(F_{A,t}, F_V, t) = \log \left[\frac{p(F_{A,t}, F_{V,t})}{p(F_{A,t}) \cdot p(F_{V,t})} \right] \gtreqless \eta, \tag{10}$$

The test function must then decide which of the hypothesis is the most likely to describe the probability density functions of the observations $F_{A,t}$ and $F_{V,t}$, by finding the threshold η that will give the best test of size α.

The mutual information is a metric evaluating the distance between a joint distribution stating the dependence of the variables and a joint distribution stating the independence between those same variables:

$$I(F_A, F_V) = \sum_{i=1}^{T-1} \sum_{j=1}^{T-1} \left[p(F_{A,i}, F_{V,j}) \log \left(\frac{p(F_{A,i}, F_{V,j})}{p(F_{A,i}) \cdot p(F_{V,j})} \right) \right]. \tag{11}$$

The link with the hypothesis test of Eq. (8) seems straightforward. Indeed, as the number of observations $F_{A,t}$ and $F_{V,t}$ grows large, the normalized log-likelihood ratio approaches its expected value and becomes equal to the mutual information between the random variable F_A and F_V [6]. The test function can then be defined as a simple evaluation of the mutual information between audio and video random variables, with respect to a threshold η'. This result differs from the approach of Fisher *et al.* in [4], where the mouth region which exhibits the largest mutual information value is assumed to have produced the speech audio signal. The formulation of the hypothesis test with a Neyman-Pearson approach allows to define a measure of confidence on the decision taken by the classifier, in the sense that the α-β trade-off is known.

Considering that two mouth regions could potentially be associated to the current audio signal and defining one hypothesis test (with associated thresholds η_1 and η_2) for each of these regions, four different cases can occur:

1. $I_1(F_A, F_{V_1}) > \eta_1$ and $I_1(F_A, F_{V_2}) < \eta_2$: speaker 1 is speaking and speaker 2 is not;
2. $I_1(F_A, F_{V_1}) < \eta_1$ and $I_1(F_A, F_{V_2}) > \eta_2$: speaker 2 is speaking and speaker 1 is not;
3. $I_1(F_A, F_{V_1}) < \eta_1$ and $I_1(F_A, F_{V_2}) < \eta_2$: none of the speaker is speaking;
4. $I_1(F_A, F_{V_1}) > \eta_1$ and $I_1(F_A, F_{V_2}) > \eta_2$: both speakers are speaking.

The experimental conditions are defined so as to eliminate the possibilities 2 and 3: the test set is composed of sequences where speakers 1 and 2 are speaking each in turn, without silent states. This allows, in the context of this preliminary work, to define the simpler following cases: if a speaker is silent, it implies that the other one is actually speaking. Notice also that a possible equality with the threshold is solved by attributing randomly a class to the random variable pair.

3.2 Hypothesis Testing for Performance Evaluation

The formulation of the previous hypothesis test gives a mean of evaluating the whole classification chain performance. Receiver Operating Characteristic (ROC) graphs allow to visualize and select classifiers based on their performance [9]. They permit to crossplot the size and power of a Neyman-Pearson test, thus to evaluate the ability of a classifier to produce good relative instance scores. Our purpose here is not to focus the evaluation on the classifier itself but on the possible gain offered by the introduction of the feature optimization step in the classification process.

To this end, two kinds of audio features are used in turn to estimate the mutual information in each mouth region: the first ones are the linear combination of the MFCCs resulting from the optimization described in sec. 2; the second ones consist simply in the mean value of these MFCCs. The results about this comparison are presented in the next section.

4 Results

4.1 Experimental Protocol

The sequence test set is composed of the eleven two-speakers sequences *g11* to *g22* [1], taken from the CUAVE database [10], where each speaker utters in turn two digit series. These sequences are shot in the NTSC standard (29.97fps, 44.1kHz stereo sound). For the purpose of the experiments, the problem has been restricted to the case where one of the speaker and only one of them is speaking in any case. Therefore, the last seconds of the video clips where the two speakers are speaking all together, as well as the silent frames - labelled as in [11] - have been discarded.

For all the sequences, the $N \times M$ mouth regions are extracted, using the face detector described in [8] (N and M varying between 30 and 60 pixels, depending on speakers' characteristics and acquisition conditions). Thus the video feature set is composed of the $N \times M \times (T-1)$ values of the optical flow norm at each pixel location (T being the number of video frames within the analyzing window, *i.e.* $T = 60$ frames). From the audio signal, 12 mel-cepstrum coefficients are computed using 30ms Hamming windows.

The optimization is done over a $2s$ temporal window, shifted by one second steps over the whole sequence to take decisions every seconds. The output of the classifier for each window is compared to the corresponding ground truth label, defined as in [11]. The test set is eventually composed of 188 test points (windows), with one audio and

[1] *g18* has been discarded as it exhibits strong noise due to the compression.

one video instances for each window. The two classes, ”speaker1” (speaker on the left of the image) and ”speaker2” (speaker on the right) are well balanced since theirs set sizes are 95 and 93 respectively.

4.2 Performance of Hypothesis Testing as a Classifier

Firstly, the ability of hypothesis testing to act as a classifier is discussed. The evaluation of the possible gain offered by using optimized audio features with respect to simpler ones is addressed in the next paragraph. Thus, only optimized audio features are put in the classifier, defined as the test function giving the best test of size α.

For binary tests, a positive and a negative class have to be defined. We assume the positive class to be the class ”speaker” for each test. More precisely, since the experimental conditions implies that there is always one speaker speaking, the positive class is the label of the mouth region where the test is performed: *i.e*, ”speaker1” for test1 (defined between the random variables F_A and F_{V_1}), and ”speaker2” for test2. Table 1 compares the power of the tests for given sizes α.

Table 1. Power of the tests for different sizes α. The thresholds η defining the corresponding decision functions are also indicated.

	Test1			Test2		
α	5%	10%	20%	5%	10%	20%
β	37.9%	81.1%	90.5%	4.3%	24.7%	89.26%
Threshold	0.41	0.25	0.16	0.55	0.45	0.25

Let us introduce now the accuracy of a test as the sum of the true positive and true negative rates divided by the total number of positive and negative instances [9]. Table 2 gives the classifier scores for the threshold corresponding to each test best accuracy: 86.7% and 85.11% for test1 and test2 respectively, obtained for thresholds $\eta_1 = 0.18$ and $\eta_2 = 0.19$.

Table 2. Detection probabilities β and false-alarm rates α for each class of each test at its best accuracy value.

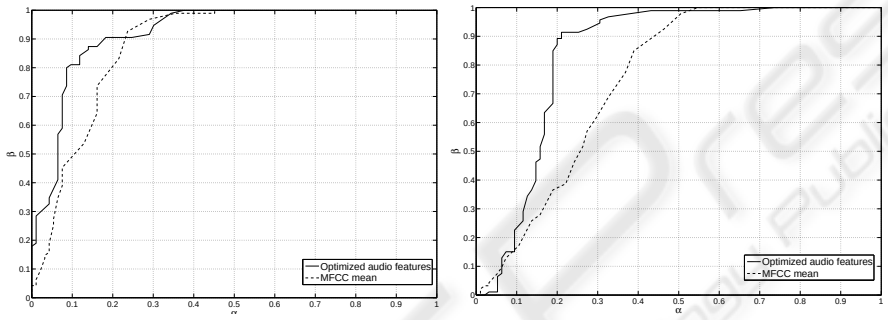
	Test1		Test2	
	Positive class	Negative class	Positive class	Negative class
β	87.4%	86.0%	91.4%	79.0%
α	14.0%	12.6%	21.0%	8.6%

These results indicate hypothesis test as a good method for assigning a speaker class to mouth regions, with a given α-β trade-off. The classifier produces better relative instance scores for test1. However, the thresholds giving the best accuracy values are about the same for the two tests. This tends to indicate that this threshold is not speaker

dependent. Further tests on larger test sets would be necessary however for a more precise analysis of the classifier capacity.

4.3 Evaluation of the Classification Chain Performance

The advantage of using optimized audio features against simple ones at the input of the classifier is now discussed. As in the previous paragraph, two tests are considered, with the positive classes being respectively the speaker 1 and the speaker 2. The ROC graphs corresponding to each test are plotted on Fig. 2. An analysis of these curves shows that the classifier fed in with the optimized audio features performs better in the conservative region of the graph (northwest region).



(a) (b)

Fig. 2. ROC graphs for tests 1 (a) and 2 (b). The detection probability for the positive class is plotted versus the false-alarm rate.

Table 3 sums up some interesting values attached to the ROC curve such as the area under the curve (AUC), or the accuracy with corresponding thresholds. Whatever the way of considering the problem, the use of the optimized audio features improved the classifier average performance, as stated by the theory in sec. 2.

Table 3. Area under the curve and accuracy with the corresponding threshold for each test.

	Test 1		Test 2	
Input features	MFCCs mean	Optimized audio features	MFCCs mean	Optimized audio features
AUC	0.88	0.92	0.75	0.84
Accuracy	$84,6\%$	$86,7\%$	$73,4\%$	$85,1\%$
Threshold	0.14	0.18	0.10	0.19

5 Conclusions

This work addresses the problem of labelling mouth regions extracted from audio-visual sequences with a given speaker class label, using both the audio and video content. The

problem is cast in a hypothesis testing framework, linked to information theory. The resulting classifier is based on the evaluation of the mutual information between the audio signal and the mouths' video features with respect to a threshold, issued from the Neyman-Pearson lemma. A confidence level can then be assigned to the classifier outputs. This approach results in the definition of an evaluation framework. The latter is not used to determine the performance of the classifier itself, but considers rather rating the whole classification process efficiency.

In particular, it is used to check whether a feature extraction step performed prior to the classification can increase the accuracy of the detection process. Optimized audio features obtained through an information theoretic feature extraction framework fed in the classifier, in turn with non-optimized audio features. Analysis tools derived from hypothesis testing, such as ROC graphs, establish eventually the performance gain offered by introducing the feature extraction step in the process.

As far as the classifier itself is concerned, more intensive tests should be performed in order to draw robust conclusions. However, preliminary remarks tend to indicate that a hypothesis-based model can be used with advantage for multimodal speaker detection.

It would also be interesting to consider in future works the cases of simultaneous silent or speaking states (cases 3 and 4 defined in sec. 3).

References

1. Hershey, J., Movellan, J.: Audio-vision: Using audio-visual synchrony to locate sounds. In: Proc. of NIPS. Volume 12., Denver, CO, USA (1999) 813–819
2. Nock, H.J., Iyengar, G., Neti, C.: Speaker localisation using audio-visual synchrony: An empirical study. In: Proceedings of the International Conference on Image and Video Retrivial (CIVR), Urbana, IL, USA (2003) 488–499
3. Butz, T., Thiran, J.P.: From error probability to information theoretic (multi-modal) signal processing. Signal Processing **85** (2005) 875–902
4. Fisher III, J.W., Darrell, T.: Speaker association with signal-level audiovisual fusion. IEEE Transactions on Multimedia **6** (2004) 406–413
5. Besson, P., Popovici, V., Vesin, J.M., Kunt, M.: Extraction of audio features specific to speech using information theory and differential evolution. EPFL-ITS Technical Report 2005-018, EPFL, Lausanne, Switzerland (2005)
6. Ihler, A.T., Fisher III, J.W., Willsky, A.S.: Nonparametric hypothesis tests for statistical dependency. IEEE Transactions on Signal Processing **52** (2004) 2234–2249
7. Moon, T.k., Stirling, W.C.: Mathematical Methods and Algorithms for Signal Processing. Prentice hall (2000)
8. Meynet, J., Popovici, V., Thiran, J.P.: Face detection with mixtures of boosted discriminant features. Technical Report 2005-35, EPFL, 1015 Ecublens (2005)
9. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Laboratories (2003)
10. Patterson, E., Gurbuz, S., Tufekci, Z., , Gowdy, J.: Cuave: a new audio-visual database for multimodal human-computer interface research. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Volume 2., Orlando, IEEE (2002) 2017–2020
11. Besson, P., Monaci, G., Vandergheynst, P., Kunt, M.: Experimental evaluation framework for speaker detection on the cuave database. Technical Report TR-ITS-2006.003, EPFL, 1015 Ecublens (2006)