IDIAP RESEARCH REPORT

Mapping Nonverbal Communication into Social Status: Automatic Recognition of Journalists and Non-journalists in Radio News

A.Vinciarelli [a] [b]

IDIAP–RR 07-33

August 2007

submitted for publication

[a] IDIAP - vincia@idiap.ch

[b] Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)

IDIAP Research Report 07-33

Mapping Nonverbal Communication into Social Status: Automatic Recognition of Journalists and Non-journalists in Radio News

A.Vinciarelli

August 2007

submitted for publication

Abstract. This work shows how features accounting for nonverbal speaking characteristics can be used to map people into predefined categories. In particular, the results of this paper show that the speakers participating in radio broadcast news can be classified into journalists and non-journalists with an accuracy higher than 80 percent. The results of the approach proposed for this task is compared with the effectiveness of 16 human assessors performing the same task. The assessors do not understand the language of the data and are thus forced to use mostly nonverbal features. The results of the comparison suggest that the assessors and the automatic system have a similar performance.

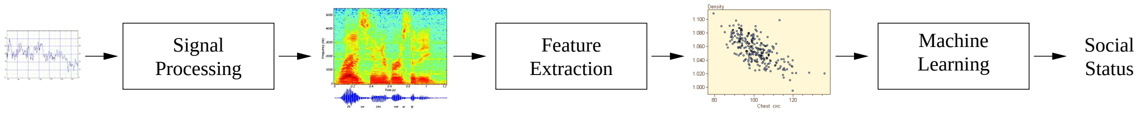

Figure 1: Approach. The figure shows the flow of the data through the system. The first stage applies Signal Processing techniques to the raw input data to track the pitch. This last is used at the feature extraction stage in order to convert the audio into vectors in an appropriate feature space. Machine Learning techniques are then applied to the vectors to perform the recognition of the social status.

1 Introduction

Sociologists use the term *status* to define the position of an individual in a given social environment [27]. Social statuses are perceived as objective by the members of the environment and are, in general, related to observable and measurable characteristics [28]. One of the simplest examples of status is the position of an individual in the organizational hierarchy of an organization. In this case, the status is clearly defined and explicitly stated. A more complex example is the status of a parent. In this case, the status characteristics are not defined explicitly, but each culture has a socially accepted idea of parenthood to which people tend to conform [28].

The nonverbal communication literature shows that people are effective in recognizing the social status by just listening to the way individuals talk, and this apply in particular to the cases where the status is identified with professional activities [12],[8]. The work presented here shows that, in the specific case of radio broadcast news, the social status can be recognized automatically using nonverbal characteristics of the way people talk. In particular, the experiments presented in this work show that the speakers can be split into journalists and non-journalists using purely acoustic features. In other words, speakers are classified as journalists and non-journalists with an accuracy of around 80 percent without taking into account what they say, but simply the way they talk. The features used in the experiments are inspired by the nonverbal communication literature and, in particular, by *vocalics*, i.e. the study of nonverbal aspects in the human voice [14][25].

The approach is illustrated in Figure 1: the first stage applies signal processing techniques to split the speech signal into voiced and non-voiced segments, the second stage extracts features accounting for nonverbal speaking characteristics, and the third one applies machine learning techniques to classify speakers into journalists and non-journalists. The results are obtained over a corpus of 686 audio clips where each one containing a single speaker intervention. The number of speaking subjects is 330 and 284 of them are represented only in the training set or only in the test set. This ensures that the system does not recognize the identity of the speakers, but their actual way of speaking.

The results of the automatic system are compared to those of human assessors performing a similar task: 16 persons were asked to classify the speaker in 30 randomly selected audio clips from the abovementioned corpus into the journalist or non-journalist category. The audio clips are in French and all assessors had no or limited understanding of such language. This ensures that they classify the clips using mostly nonverbal characteristics and not what is said. The mother tongues of the assessors include English, Serbian, Hindi, Chinese, Farsi and Arabic. This significantly limits the possibility that the assessors use their native language to understand even partially what the speakers say. Overall, the performance of the human assessors is around 80 percent and this seems to suggest that the system is as effective as humans in recognizing the social status through purely nonverbal features. However, since the assessment is time consuming, the test had to be performed on a subset of the whole corpus, thus the comparison is only indicative.

The recognition of social status can be useful in several applications: browsing systems can be enhanced by showing the social status of speakers in audio or video recordings, indexing and retrieval systems can include the status information in the data features in order to enrich the spectrum of queries that can be addressed, and summarization systems can use the status to select only interventions that are likely to provide useful information. In more general terms, the attempt to extract information based on noverbal communication is a step towards social aware systems, which are sensitive to the same kind of social signals which humans use to communicate with each other [5],[21]. To our knowledge, no other works presented in the literature have used to classify individuals into predefined categories and this is the main novelty of this work.

The rest of this paper is organized as follows: Section 2 presents a survey of previous work, Section 3 describes the extraction of nonverbal features, Section 4 introduces the recognition approach, Section 5 shows experiments and results, and Section 6 draws some conclusions.

2 Previous Works

This section presents a survey of works where nonverbal characteristics of speech are used to extract information from the content of multimedia data. Nonverbal behaviors include body position, facial expression, gestures, etc., but this survey focuses on speech because this is the source of nonverbal features used in this work. The papers presented in this section can be grouped into three major areas: the first is the recognition of the affective or emotional state of people, the second is the extraction of information about social interactions, and the third concerns the detection of the so-called *speaking mode*, i.e. a set of prosodic features used to improve the performance of speech recognition systems.

Emotion recognition approaches are typically grouped under the collective name as that of *affective computing* and involve data as diverse as videos, physiological signals and eye gaze (see [22] for an extensive monograph). This quick survey focuses on techniques using audio. Some works address the problem of emotion recognition independently of an application [4],[17],[34]. The work in [4] tries to detect emotions through the *pitch*, i.e. the oscillation frequency of the vocal folds (see Section 3), the experiments in [17] concern spoken dialogues where the detection of the affective state provides clues about the development of the discussion, and the work in [34] shows the advantages of combining features extracted from both audio and video channels of the same recording. In the multimedia community, the emotions are typically used as a low level feature to infer higher level information [11],[10],[3]. Emotions have been used to identify the most important moments of sport videos [10] or movies [3], as well as indexes in retrieval systems [11].

The extraction of information about social interactions from audio is not as established as the affective computing and most of the approaches proposed so far are still at a relatively early stage. Several pioneering works have been dedicated to the extraction of *social signals*, i.e. nonverbal behaviors that human use in social interactions [1]. Audio features aiming at capturing social signals have been shown to be effective in predicting with high accuracy (more than 75 percent) the outcome of job interviews and salary negotiations [6], and the centrality of individuals in professional social networks [7],[20]. Other works address the problem of finding the most dominant person in meetings using not only audio, but also video recordings and the output of other sensors [13],[26]. Meeting recordings are analyzed also in [33] to map the behavior of people into a predefined set of sociological models, and in [15] to assess the satisfaction of people involved in group discussions. Recent works use the output of speaker clustering techniques to extract social networks allowing one to recognize the roles of people talking in broadcast news [29] and to segment news bulletins into stories [30]. Similar techniques, but based on video, have been used to detect the main characters in movies [31].

The last group of papers includes works where the analyisis of nonverbal speaking characteristics aims at improving the performance of speech recognition systems. Some works recognize the disfluencies (hesitations, pauses, repetitions, etc.) to avoid recognition errors [18],[16],[32]. Other approaches, specifically targeting the recognition in broadcast news, try to detect the *speaking mode*, i.e. try to distinguish between different ways of speaking (fluent and continuous, slow and hesitant, etc.) to adapt the recognition system [19],[9]. One approach considers the speaking mode as a hidden variable [19], while the other models the influence of the speaking rate on the word pronunciations [9].

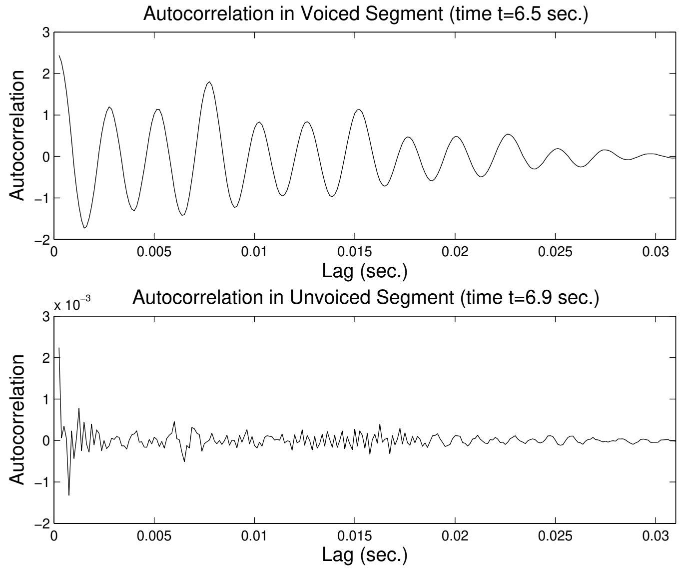

Figure 2: Autocorrelation function. The upper plot shows the autocorrelation function extracted from a voiced segment. The period is roughly 2.5 ms which corresponds to a frequency of 400 Hz. The lower plot is extracted from an unvoiced segment and no peridicity is observed.

3 Nonverbal Features Extraction

This section presents the nonverbal features extracted from the voice of the speakers. The extraction process includes two steps: the first is the estimation of the pitch and the second is the estimation of the relative entropy of voiced and unvoiced segment length distributions. The next sections show the two steps in detail.

3.1 Pitch Estimation

When humans produce voiced sounds, they push air from the lungs to the vocal tract. At the beginning of the vocal tract, the air passes through the *glottis* where the *vocal folds* oscillate like the cords of a guitar. The oscillation frequency is called *pitch* and it is the characteristic that alone contributes more than anything else to the quality of a voice [23]. For this reason, the automatic estimation of the pitch has been investigated extensively since the early times of speech processing and the literature offers a wide spectrum of methods for this task [24]. In this work, the pitch is used to split speech data into *voiced* and *unvoiced* segments, i.e. into segments where the vocal folds oscillate and do not oscillate respectively. Unvoiced segments do not correspond only to silences, where there is no emission of sound at all, but also to a certain number of elementary sounds, the so-called *phonemes*, that are used in a given language to compose all possible words. An example of unvoiced phoneme is the sound $/s/$ at the beginning of the words *start* and *sale*.

Since the goal of the pitch estimation is the discrimination between voiced and unvoiced segments, it is not necessary to have an accurate pitch estimate, but simply one that is accurate enough to distinguish between the two above conditions. For this reason, the pitch estimation (or pitch tracking) technique used in this work is relatively simple. The approach includes three steps: the first is the estimation of the pitch using the autocorrelation function, the second is the estimation of the pitch using the Fourier Transform, and the third is the averaging of the two estimates. The use of two pitch estimation techniques aims at making the method more robust.

The first step of the process is the extraction of the *autocorrelation function*:

$$R_n[k] = \sum_{m=-\infty}^{\infty} s[m]w[n-m]s[m+k]w[n-m-k] \tag{1}$$

where $s[n]$ is the speech signal, $w[l]$ is the analysis window, n is an integer number ranging between $-\infty$ and ∞, and k is the lag. In general, the analysis window is different from zero only in an interval of finite length N, thus the autocorrelation is zero whenever the lag k is higher than N. In fact, when $k > N$ there are no values of m where both $w[n-m]$ and $w[n-m-k]$ are different from zero. In this work, $w[n]$ is a rectangular window:

$$w[n-m] = \begin{cases} 1 & 0 \leq m < N \\ 0 & otherwise. \end{cases} \tag{2}$$

The main property of the autocorrelation function is that if $s[n]$ is periodic, then $R_n[k]$ is periodic with the same period. This means that the autocorrelation function can be used to detect the dominant frequency, if any, in a signal $s[n]$. In the case of the speech, the dominant frequency is the pitch (in voiced segments), thus the autocorrelation can provide a measure, although noisy, of the pitch.

Figure 2 shows the autocorrelation function extracted from one voiced (upper plot) and one unvoiced (lower plot) segment. While the upper plot is clearly periodic, the second does not show any significant structure. The reason is that voice segments have a dominant frequency, while unvoiced segments do not. The amplitude of the autocorrelation decreases with the lag because the intersection between $w[n-m]$ and $w[n-m-k]$ becomes shorter, thus less and less terms are different from zero in the sum of Equation 1.

The pitch can be estimated by measuring the average time distance between two consecutive peaks of $R_n[k]$, i.e. the period of the autocorrelation. The inverse of such value is the pitch estimate. In our experiments, the value of N is 256, and the pitch estimation is peformed at regular time steps 128 samples appart. Since the speech signal is sampled at 8000 Hz, 256 samples correspond to roughly 30 ms. This is the minimum time required to change the configuration of the vocal tract, i.e. to change significantly the properties of the voice. In other words, the value of N corresponds to a time interval which is too short to observe significant changes in the voice properties.

The second step in the pitch tracking process is based on the Fourier Transform (FT) [24]. The FT is extracted from from signal segments of length N and at regular time steps of 128 samples. In other words, the FT is extracted at the same instants and from the same segments from which the autocorrelation is extracted. The extraction of the FT results into the so-called *spectrogram* (see Figure 3) showing the evolution of the spectral properties over time. The lower plot of Figure 3 shows one column of the spectrogram, i.e. the power spectrum extracted at the same instant when the autocorrelation of Figure 2 is extracted. The frequence with the highest energy is around 400 Hz, which corresponds to a period of 0.0025 *sec.* which is roughly the the distance between two peaks of the autocorrelation function in Figure 2 (upper plot).

The third step of the pitch tracking process is the averaging of the two pitch estimates. If $p_a[t]$ and $p_s[t]$ are the pitch estimates extracted using the autocorrelation and the spectrogram respectively, then a robust estimate $p[t]$ of the pitch can be obtained by simply averaging the two:

$$p[t] = \frac{1}{2}(p_a[t] + p_s[t]) \tag{3}$$

The value of $p[t]$ is plotted in Figure 4 and shows how it tends to have higher values in correspondence of the non-voiced segments (the peaks of the plot). The reason is that in such segments high frequency noise tends to become dominant. The average value $\bar{p}$ of the pitch estimate in a given recording can be used as a threshold to discriminate between voiced and unvoiced frames. The horizontal line in Figure 4 corresponds to the threshold and it shows how voiced and unvoiced segments can be discriminated.

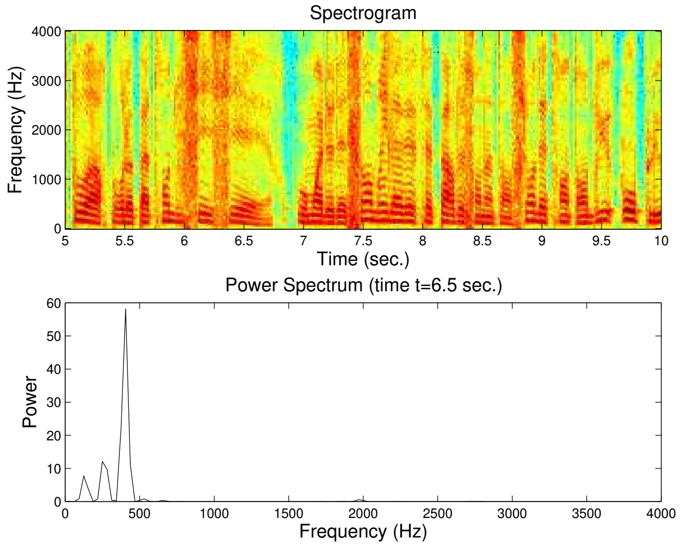

Figure 3: Spectrogram. The upper plot shows the spectrogram for five seconds of a clip. Darker areas correspond to higher energy frequencies, the parallel bands of the voiced segments are the so-called formants, i.e. integer multiples of the pitch. The lower plot shows a column of the spectrogram. The peaks correspond to the highest energy frequencies.

3.2 Nonverbal Features

After the pitch estimation, the data is split into voiced and non-voiced segments. The distribution of the segment lengths (both voiced and unvoiced) can be estimated by simply counting the number of times a given length is represented. The value of the pitch is estimated at regular time steps and the length of the intervals can be only a multiple of the length of a single step. In other words, the length of voiced and unvoiced segments is quantized and the resulting distributions are discrete. The length distribution of the voiced interval lengths is $p_v(l)$, while the one for the unvoiced segments is $p_u(l)$. In both cases, the distribution enables one to estimate the relative entropy:

$$H_i = \frac{-\sum_{l \in \mathcal{L}} p_i(l) \log p_i(l)}{\log |\mathcal{L}|} \tag{4}$$

where i is u or v, $\mathcal{L}$ is the set of the lengths represented in a given speech segment, and $|\mathcal{L}|$ is the cardinality of $\mathcal{L}$. The relative entropy is bounded between 0 and 1 and it accounts for the variation in the interval lengths: the closer H_i is to 1, the more the distribution is uniform and the variability is high.

Each speech segment can be represented with a vector $\vec{y}$ where the components are H_v and H_u. The vectors can be decorrelated using the Principal Component Analysis (PCA) and Figure 5 shows the resulting vectors $\vec{x}$ for the audio clips used in the experiments of this work (see Section 5 for more details). Despite the overlap, journalists and non-journalists seem to form two separate classes. The variability in the length of voiced and non-voiced segments seems thus to capture the status of the speakers. This is in agreement with the nonverbal communication theory because skilled speakers, i.e. speakers that are more effective in conveying their message, tend to have higher variations in nonverbal characteristics than non skilled speakers [14],[25]. Since journalists are more used to speaking in front of the microphones they are likely to have, on average, higher variation in voiced/unvoiced segment lengths and this affects the distribution of the points in the space of Figure 5.

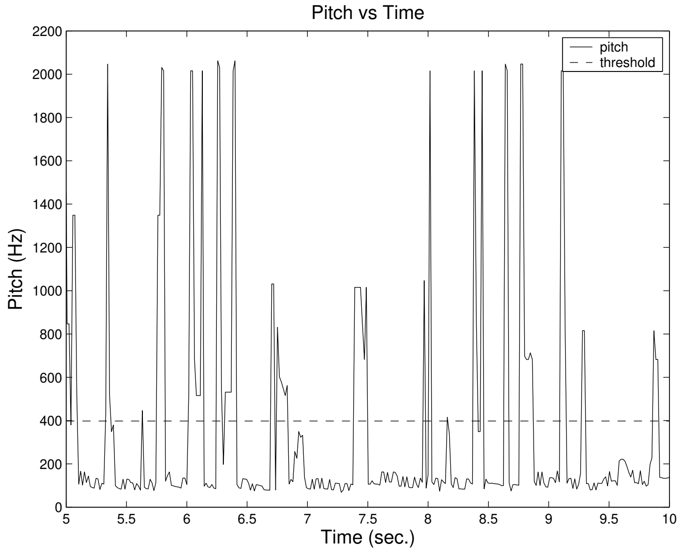

Figure 4: Pitch. The plot shows the evolution of the pitch estimate along the time. The regions above the threshold are assumed to correspond to unvoiced segments.

The length of the intervals for both voiced and unvoiced segments ranges between 0.1 and 0.5 seconds. Since nobody can control phenomena taking place at such a small temporal scale, the length of the intervals, and correspondently the two features described above, are the result of an unconscious process. This is important because, in general, the least conscious nonverbal behaviors carry the most reliable information. The reason is that such behaviors cannot be simulated, thus cannot be used to lie [14],[25].

4 Statistical Foundations

The social status recognition task can be performed by finding the status s^* which maximizes the a-posteriori probability $p(s|\vec{x})$:

$$s^* = \arg\max_{s \in \mathcal{S}} p(s|\vec{x}) \tag{5}$$

where s is the social status, $\vec{x}$ is the vector representing a given speech segment (see Section 3), $\mathcal{S}$ is the set of all possible social statuses and $p(s|\vec{x})$ is the a-posteriori probability of the social status. By applying Bayes theorem, the above equation can be rewritten as follows:

$$s^* = \arg\max_{s \in \mathcal{S}} p(\vec{x}|s)p(s) \tag{6}$$

where $p(\vec{x}|s)$ is the likelihood of the vector $\vec{x}$ given the model corresponding to status s and $p(s)$ is the a-priori probability of status s, i.e. the probability of observing s.

The problem is how we can estimate $p(s)$ and $p(\vec{x}|s)$. The a-priori probability can be estimated by simply using the fraction of training data which corresponds to a given status. The likelihood can be estimated using a multivariate Gaussian distribution [2]:

$$\mathcal{N}(\vec{x}|\vec{\mu}_s, \Sigma_s) = \frac{1}{(2\pi^D \Sigma_s)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_s)^T \Sigma_s^{-1}(\vec{x}-\vec{\mu}_s)} \tag{7}$$

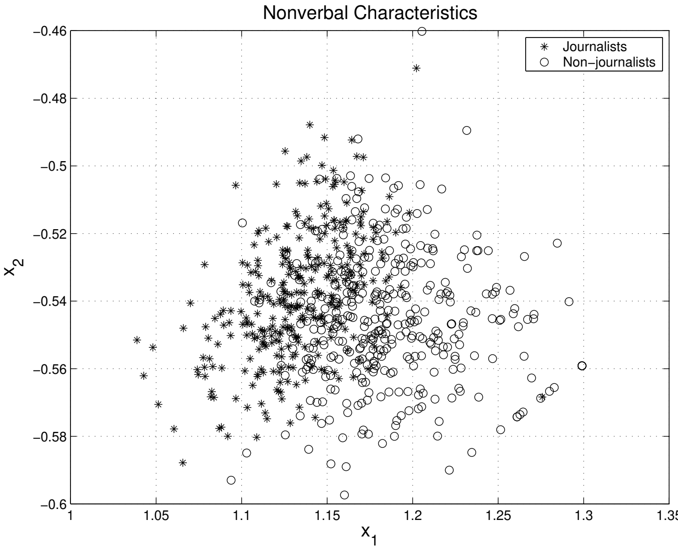

Figure 5: Scatter Plot. Each point account for a single clip. Journalists and non-journalists are plotted using a different symbol to show that they tend to form separate classes.

where D is the dimensionality of the data. Means $\vec{\mu}$ and covariance matrices Σ can be estimated by maximizing the likelihood of the model over a training set $\mathcal{X} = \{\vec{x}^{(1)}, \ldots, \vec{x}^{(M)}\}$.

In our experiments, the vectors $\vec{x}$ are the projections of the feature vectors described in Section 3 onto the Principal Components extracted from the training set [2]. This means the data is decorrelated and the covariance matrices are diagonal. In this case, the multivariate Gaussian distributions are products of Gaussians:

$$\mathcal{N}(\vec{x}|\vec{\mu}_s, \Sigma_s) = \prod_{i=1}^{D} \frac{1}{(2\pi\sigma_{si}^2)^{\frac{1}{2}}} e^{-\frac{(x_i-\mu_{si})^2}{2\sigma_{si}^2}} \tag{8}$$

where μ_{si} and σ_{si} are mean and variance of the i^{th} component in model s, and x_i is the i^{th} component of $\vec{x}$. The value of the different means and variances can be estimated by using sample means and sample variances of the components of the vectors in the training set $\mathcal{X}$:

$$\mu_{si} = \frac{1}{N_s} \sum_{n=1}^{N_s} x_i^{(n)} \tag{9}$$

$$\sigma_{si} = \frac{1}{N_s} \sum_{n=1}^{N_s} (x_i^{(n)} - \mu_{si})^2 \tag{10}$$

where N_s is the number of samples of status s in the training set.

5 Experiments and Results

The next sections present in detail the data used in this work and the results of the experiments.

5.1 Data and Experimental Protocol

The experiments of this work have been performed over a corpus of 686 audio clips extracted from 96 news bulletins broadcast by Radio Suisse Romande, the French speaking Swiss national broadcasting service, in February 2005. Each clip corresponds to an intervention of a single speaker. The total duration of the clips is 7 hours and 10 minutes, the number of journalist clips is 313 and the number of non-journalist clips is 373, corresponding to 45.6 percent and 54.4 percent of the data respectively.

The corpus was split into two partitions containing respectively 377 and 311 clips. The first partition contains clips of the first 15 days of February 2005, the second partition contains clips of the remaining 13 days of the same month. When the first partition is used as a training set, the second one is used as test set and vice-versa. This makes it possible to perform experiments over the whole corpus while still preserving the necessary separation between training and test set. All the results presented in this work are the weighted average of the performances obtained over the two different partitions.

The number of individuals in the set of non-journalist clips was 234 and only 7 people appear in both dataset partitions. This means that the identity has a negligible effect on the performance of the automatic system so the recognition is based on the actual way of speaking. In the case of the journalists, the number of identities is 96 and 39 individuals appear in both dataset partitions. This means that the effect of identity is more important than in the case of the non-journalists. The number of audio clips per identity follows a Zipf Law:

$$N_n \sim \frac{1}{n} \tag{11}$$

where N_n is the number of identities represented n times. Around one third of the identities are represented only once ($N_1 = 112$) and 60 percent of the identities are represented less than four times ($N_2 = 57$ and $N_3 = 35$). This means that in most cases the identity plays a minor role or no role at all in the recognition of the social status.

5.2 Automatic Recognition Results

The overall recognition rate of the system based on the approach depicted in Figure 1, i.e. the percentage of clips classified correctly, is 81.0 percent. The recognition rate for journalists and non-journalists is 85.9 and 76.9 percent respectively (see Section 5.1 for details about the experimental protocol). The performance can be compared with the recognition rate of two simple baseline systems. The first system simply apply the *highest a-priori probability rule*, i.e. it assigns each clip to the most probable class. In the case of our data, the most probable class is the non-journalist one and the performance when assigning all clips to such class is 54.4 percent. The second simple baseline guesses randomly the class of each clip. The probability of guessing class s is given by the a-priori probability $p(s)$ of s. In this case, the recognition rate is 50.4 percent. The difference between the performance of the system described in the previous part of the paper and the two simple baselines described above is statistically significant. This means that its performance is not due to simple chance or to an unbalanced distribution of classes in the test set.

The plot in Figure 6 shows how the performance changes when only the first t seconds of the test clips are retained to extract the features. When the system uses only the first 4 seconds of the clips, the recognition rate is 54.5 percent, similar to the performance obtained through a random guess or by applying the highest a-priori probability rule. The curve increases roughly linearly until 45 seconds, then it converges to the overall performance of 81 percent. The plateau is probably due to the fact that the number of clips longer than $45-50$ seconds is small, then the impact on the overall performance tends to decrease as the number of retained seconds increase. However, the plot seems to suggest that longer clips are easier to classify and this is not surprising: the reason is that the features are based on probability distributions estimated using the speech data (see Section 3) and longer clips allow more reliable estimations.

The effect of the clip length is probably the explanation of the recognition rate difference between journalists and non-journalists. In fact, non-journalist clips are shorter, on average, than journalist clips. Moreover, for the journalists there are more individuals represented in both training and test sets, so the identity has a higher influence than in the case of the non-journalists. However, some

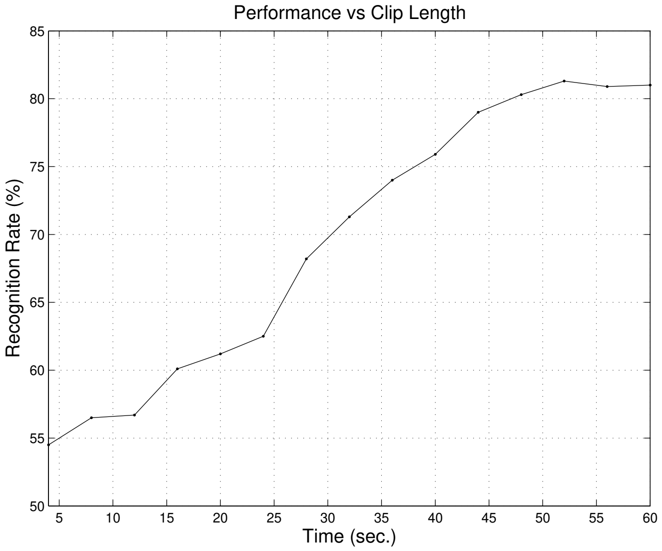

Figure 6: Recognition rate vs time. The plot shows the recognition rate as a function of the number of seconds preserved in the test clip.

misclassifications are due to the intrinsic ambiguity of the data: some non-journalist speakers, e.g. politicians and actors that often appear in the media, use the same nonverbal techniques as the journalists and, at the same time, some journalists are not as effective as their colleagues in delivering their message.

To our knowledge, no other systems performing a similar task have been presented in the literature. This makes it difficult to say whether the performance of the system is satisfactory or not. For this reason, the next section proposes the results of a test where human assessors are asked to perform the same recognition task as the system. The results will provide a term of comparison.

5.3 Test with Human Assessors

The performance of the automatic system has been compared with the results obtained by 16 human assessors on a similar task. A set of 30 audio clips were randomly selected from the data corpus. In this set, 17 clips correspond to journalists and 13 to non-journalists. The length of the clips ranges from 3.5 to 75 s and it reproduces roughly the length distribution of the data corpus.

The human assessors have listened to the clips and have assigned each one of them to one of the two classes. In order to reduce as much as possible the influence of the content, the assessors do not speak the language of the clips (French) and their mother tongues include English (2 persons), Hindi (5 persons), Chinese (6 persons), Farsi (1 person), Serbian (1 person) and Arab (1 person). The group of the assessors includes 5 women and 11 men.

The total number of judgments made by the assessors is 480 and their overall performance, i.e. the fraction of correct judgments, is 82.3 percent. The women have an overall performance of 88 percent (on average 26.4 correct judgments out of 30), while the men have an overall performance of 79.0 percent (on average 23.7 correct judgments out of 30). On average, each clip has been recognized correctly by 13.2 assessors, but there are two ambiguous clips, recognized by only 2 and 4 assessors respectively, that reduce significantly the average. Without taking into account such clips, the average number of correct classifications per clip is 13.9. The correlation coefficient between the clip length and the number of correct classifications is 0.44. This means that the length does not affect significantly the performance of the assessors and the results are roughly the same while being independent of the length of time the speaker spoke. This is an important difference with respect to the automatic system which is sensitive to the amount of time retained to extract the features (see Figure 6).

The comparison is only indicative because the test set is not the same for the automatic system and for the human assessors. The reason is that the assessors should listen to the whole test set (more than seven hours) and this is not possible for practical reasons. However, the results seem to suggest that the humans have a performance similar to the algorithm when they rely on nonverbal communication. The performance of the automatic system over the 30 clips used during the assessment is 73.3 percent. The difference with respect to the performance of the assessors is not statistically significant due to the small number of test samples.

6 Conclusion

This paper has presented experiments where features inspired by the nonverbal communication literature have been used to discriminate between journalists and non-journalists in radio broadcast news. The results show that the recognition rate of the automatic system is around 80 percent although the classifier is a simple multivariate Gaussian. The recognition rate of the system has been compared with the results obtained by 16 human assessors performing a similar task. The assessors do not understand the language of the data (French) and their judgments are based mostly on nonverbal aspects. The average recognition rate of the human assessors is around 80 percent and this seems to suggest that the automatic system is not far from the human performance. However, the assessors used a subset of the data corpus used during the experiments, so the comparison is only indicative. The performance of the system over the subset used by the human assessors is around 70 percent, but the number of audio clips (30) is too small to conclude that the difference is statistically significant.

The most important novelty of this work is that the content of the audio clips, i.e. what the people say, has not been taken into account. The classification has been performed using only two features which account for micro-characteristics of the speech; the time intervals taken into account are less than 0.5 seconds, so it cannot be consciously controlled by the speakers. This seems to suggest that nonverbal characteristics can not only be detected automatically, but also used to infer higher level information, in this case the status of the speakers, potentially difficult to extract by other means. To our knowledge, this is the first work which uses nonverbal features to map speakers into predefined categories.

The main advantage of nonverbal features is that they require relatively simple technologies, like the pitch tracking technique used in this work, and allows one to avoid the automatic transcription of the speech that is typically needed in content based approaches, i.e. in techniques that try to infer high level information using what people say. Moreover, nonverbal features are language independent and, to a certain extent, culture-independent. Thus they can cope with multilingual and multicultural data sources.

The main limit of this work is that the people in broadcast news are less spontaneous than in data like meeting recordings or home videos. For this reason, future work will focus on different kinds of data where nonverbal features can face more difficult problems. Moreover, one of the most important aspects of nonverbal features is that it plays a major role in social interactions, thus approaches similar to the one presented here should be used to extract information from the exchanges between different individuals rather than from a single person at a time.

Acknowledgements. This work was supported by the European Union under the integrated projects AMIDA, Augmented Multi-party Interaction with Distance Access, contract number IST-033812, as well as the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The author gratefully thanks the EU and Switzerland for their financial support, and all project partners for a fruitful collaboration. More information about AMIDA and IM2 is available from the project web sites www.amiproject.org and www.im2.ch. The author wishes to thank Katayoun Farrahi, Sriram Ganapathy, Phil Garner, Weina Ge, Hayley Hung, Dinesh Jayagopi, Jie Luo, Wanjun Jin, Eileen Lew, Weifeng Li, Joel Pinto, Samuel Thomas, Tamara Tosic, Muhammad Ullah, Deepu Vijayasanen and Jin Yao. Hayley Hung is gratefully acknowledged for commenting on the draft.

References

[1] N. Ambady, F. Bernieri, and J. Richeson. Towards a histology of social behavior: judgmental accuracy from thin slices of behavior. In M.P. Zanna, editor, *Advances in Experimental Social Psychology*, pages 201–272. 2000.

[2] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] C.H. Chan and G.J.F. Jones. Affect-based indexing and retrieval of films. In *proceedings of ACM International Conference on Multimedia*, pages 427–430, 2005.

[4] N.D. Cook, T.X. Fujisawa, and K. Takami. Evaluation of the affective valence of speech using pitch substructure. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):142–151, 2006.

[5] J. Crowley. Social perception. *ACM Queue*, 4(6):34–43, 2006.

[6] J. Curhan and A. Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first five minutes. *Journal of Applied Psychology (to appear)*, 2007.

[7] N. Eagle and A. Pentland. Reality mining: sensigng complex social signals. *Journal of Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[8] D.S. Ellis. Speech and social status in America. *Social Forces*, 45:431–451, 1967.

[9] E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word predictability on conversational pronunciations. *Speech Communication*, 29(2-4):137–158, 1999.

[10] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Transactions on Multimedia*, 7(6):1114–1122, 2005.

[11] A. Hanjalic and L.Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[12] L.S. Harms. Listener judgments of status cues in speech. *Quarterly Journal in Speech*, 47:164–168, 1961.

[13] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classifiy the most dominant person in a group meeting. In *proceedings of ACM International Conference on Multimedia*, 2007.

[14] M.L. Knapp and J.A. Hall. *Nonverbal Communication in Interaction*. Harcourt College Publisher, 1996.

[15] O. Kulyk, J. Wang, and J. Terken. Real time feedback on nonverbal behavior to enhance social dynamics in samll group meetings. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction*, pages 150–161. 2005.

[16] M. Lease, M. Johnson, and E. Charniak. Recognizing disfluencies in conversational speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1566–1573, 2006.

[17] C.M. Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.

[18] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1526–1540, 2006.

[19] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciations via a language dependent hidden speaking mode. In *proceedings of International Conference on Spoken Language Processing*, 1996.

[20] A. Pentland. Social dynamics: signals and behavior. In *proceedings of International Conference on Developmental Learning*, 2004.

[21] A. Pentland. Socially aware computation and communication. *IEEE Computer*, 38(3):33–40, 2005.

[22] R.W. Picard. *Affective Computing*. MIT Press, 2000.

[23] J.O. Pickles. *Introduction to the Pshysiology of Hearing*. academic Press, 1988.

[24] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[25] V.P. Richmond and J.C. McCroskey. *Nonverbal Behavior in Interpersonal Relations*. Allyn and Bacon, 1995.

[26] R. Rienks and D. Heylen. Dominance detection in meetings using easily obtainable features. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction*, pages 76–86. 2005.

[27] J. Scott and G. Marshall. *Oxford Dictionary of Sociology*. Oxford University Press, 2007.

[28] H.L. Tischler. *Introduction to Sociology*. Thomson Wadsworth Publishing, 2006.

[29] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(9), 2007.

[30] A. Vinciarelli and S. Favre. Broadcast news story segmentation using Social Network Analysis and Hidden Markov Models. In *proceedings of ACM International Conference on Multimedia*, 2007.

[31] C.Y. Weng, W.T. Chu, and J.L. Wu. Movie analysis based on roles social network. In *proceedings of IEEE International Conference on Multimedia and Expo*, pages 1403–1406, 2007.

[32] J.-F. Yeh and C.-H. Wu. Edit disfluency detection and correction using a cleanup language model and an alignment model. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1574–1583, 2006.

[33] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *proceedings of International Conference on Mutlimodal Interfaces*, pages 47–71, 2006.

[34] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, and S. Levinson. Audio-visual affect recognition. *IEEE Transactions on Multimedia*, 9(2):424–428, 2007.