INFERRING SOCIAL RELATIONSHIPS IN A PHONE CALL FROM A SINGLE PARTY'S SPEECH

Sree Harsha Yella[1,2*], *Xavier Anguera*[1] *and Jordi Luque*[1]

[1]Telefonica Research, Barcelona, Spain
[2]Idiap Research Institute, CH-1920 Martigny, Switzerland
sree.yella@idiap.ch, xanguera@tid.es, jls@tid.es

ABSTRACT

People usually speak differently depending on who they talk to. Based on this hypothesis, in this paper we propose an automatic method to detect the social relationship between two people based solely on a set of acoustic and conversational characteristics. We argue that changes in these features of an individual reflect the social relationship with the other person. To infer relationship we only require the speech of one of the conversation partners and the interaction patterns between both speakers. We validate the proposed system using a real-life telephone database with calls made by several speakers to close family members and to their partners. We trained a classifier using a boosting algorithm on a set of conversational and acoustic features and use it to classify calls according to the social relationship between both speakers. Tests performed on models trained on single speaker's data show that for most people such prediction is feasible. We also show that these characteristics generalize quite well across speakers, achieving around 75% accuracy when both sets of features are combined.

Index Terms— Speech communication, social relationship, interpersonal stance, conversational speech, prosodic features, turn-taking features

1. INTRODUCTION

The phenomenon one could refer to as *speech social role adaptation* involves the changes in the speech characteristics of an individual depending on the people being spoken to. Take as an example the differences that can be perceived when a person speaks to his parents versus when he speaks to his boss or to his son. While such differences might be clear by analyzing the content of the conversations, in some cases they are also perceivable at the acoustics level (pitch, acoustic range, etc.) and at the conversational level (speech pacing, dialogue interactions, etc.).

In this paper we are interested in inferring social relationships by means of the speech recorded from spontaneous telephone conversations between several subjects and people in their social network. In particular, given a set of calls from the same person to different people, we aim to classify such calls according to the relationship of the caller with these people. In doing so we only use the recorded speech of the caller and conversational information extracted from the call. No audio is used from the person on the other side of the phone line. This constraint allows us to model the changes in the speaker's voice or communication manner that have to do with his social relationship with the other person, and avoid the system from using any information of the callee. This might be interesting for applications where, due to privacy reasons, only the voice of the caller (who accepts being recorded) should be exposed to analysis systems, or to automatically develop a map of relationships of the user with his social network by analyzing how he interacts with the different people he talks to on the phone.

While the phenomenon of role adaptation depending on who the interlocutor is can be easily perceived when observing people talking to each other, very few studies have analyzed this phenomenon in the literature. Within the sociology area, in [1, 2] they talk about the interactions between the style of speech and the social relationship between individuals. Within the speech area, we find the work of Campbell et al. [3, 4, 5] using acoustic cues, and the work of Stark et al. [6] using textual cues. In [3, 4] the authors assert that certain acoustic features (which they call *voice quality* features) vary according to the social relationship between speakers. To prove this, a set of acoustic features are analyzed in order to show how some of them present a bigger variance when the relationship between the two speakers in a conversation changes. Later on, in [5] they analyze how laughter also changes depending on who we are laughing with. Unlike in [4], in our study we not only consider acoustic features, but also conversational features. In addition, for our study we use a database recorded in an unconstrained setting, whenever and wherever the test subjects want to make a phone call. Furthermore, we move further from just the analysis of changes and also propose and test a classification method to automatically identify the social relationship between both persons in a conversation. In [6], a similar task to ours is performed but focusing on the content of the conversation. They use the output of an automatic speech recognition system to obtain lexical information on a long-term phone calls recordings dataset, with the aim to measure and detect cognitive decay and depression on elder people.

The dataset used in this work was collected by volunteer

*At the time of this work, S. Yella was visiting Telefonica Research

subjects that were allowed to make free calls to people in their social network in exchange for, among other information, the recorded audio and the information of whom they were talking to. After selecting a subset of these subjects we extracted acoustic features from the caller's voice and conversational features derived from the turn-taking patterns from the two-side conversation. We then used a boosting-based classifier in a two social roles classification task. Aiming to assess the speaker independence of our approach, we tested the classification performance both at the speaker level (train and test on different calls from the same speaker) and at global level by building a global model with data from all subjects. Results show that not all speakers present a comparable variability in their speech according to whom they are talking to. Nonetheless, when building a general model, we found that it can generalize and is able to identify with an accuracy of around 75% the social relationship between two speakers.

2. DATASET DESCRIPTION

The dataset employed in this paper consists of Spanish two-side telephone conversations obtained from the CallNotes dataset [7]. The dataset contains regular phone calls between 62 users and people from their regular social network. Each user participating in the database creation installed a voice-over-ip application in their phone and was able to make unlimited phone calls to regular phones or cellphones in exchange for these being recorded. Before connecting each phone call a prompt was played to the callee to inform them that they would be recorded. Once a phone call finished, the participant was asked to login to a website and fill out a questionnaire about the call. Among other things, this questionnaire included his social relationship with the callee. The possible relationships the caller could select from were: partner(girlfriend/boyfriend or wife/husband), family(understood as parents or other relatives), friends, co-worker, business (meaning calling a restaurant, the doctor, etc.). Over all, the participants made 796 calls but only 305 of them were annotated a posteriori by the user. For each call, we have two individual channels that contain the audio from each side of the conversation, the caller and the callee respectively.

In order to build classification models, we selected the users that annotated at least 3 calls for each of two or more different relationships. Most of the annotated calls belong to the family and partner categories and many users only used the application to communicate with people from a single social relationship type. We were able to extract 6 users that labelled at least 3 calls on two categories (partner and family members). All the experiments reported in this work are performed on the calls from these users. On the whole there are 49 calls to a partner and 30 calls to other family members. Whilst we are aware that the final database is limited and it contains two relationship classes that might a priori be considered very similar, as we will see in the results section, differences can still be appreciated between these two classes for most users, and a general model to classify between these two social relationships can be effectively built.

3. SOCIAL RELATIONSHIP CLASSIFICATION

3.1. Data Preprocessing

In this work we use only the audio from the caller and the conversational information extracted from the interaction between both speakers. To obtain these, we automatically preprocess the recording of the call. Initially, a two-channel recording is available with the speech from each of the persons in the call. Each of the two channels is processed using a Speech Activity Detection (SAD) algorithm, based on short-term spectral energy features. From these features we train a GMM model for each of two classes (speech and non-speech) independently for each channel in the phone call. Initially, the non-speech class is bootstrapped using the 30% of lowest energy frames, and the rest is used for modeling the speech class. It is expected that the non-speech model will model both the non-speech frames and those frames with very low energy that correspond to cross-talk interferences across channels. The final SAD hypothesis is obtained by iterative re-estimation and re-alignment steps of these models until convergence. Acoustic feature extraction is performed on each of the isolated speech segments identified in the channel corresponding to the caller (provided they are longer than 1 second) while the interaction patterns between both speakers are used as conversational features.

3.2. Feature extraction

3.2.1. Acoustic Feature Extraction

The acoustic features used in this study are extracted using the openSMILE toolkit [8]. We have used one of the standard feature extractor setups distributed with the toolkit which was used as a baseline for Interspeech 2010 paralinguistic challenge [9]. The features are extracted from a set of low-level feature descriptors such as MFCC, F0, LSP and Intensity. The toolkit then computes various regression coefficients over the feature envelopes of each of the low-level feature descriptors. It also computes several functionals such as minimum and maximum values, percentiles, and quartiles over the feature envelopes. A summary of the features extracted is presented in Table 1 and can also be found in [9]. One 1582-dimension feature vector is extracted for each caller segment greater than one second in duration.

3.2.2. Conversational Feature Extraction

Conversational features are extracted from the speaker turns in a call obtained from the SAD output on each channel in a telephone conversation. Several works have previously explored these features to characterize the role of participants in multi-party conversations [10], and also to characterize a conversation on the whole [11]. These works have shown that the turn-taking patterns of an individual speaker can carry useful information in identifying the speaker [12] and also it is possible to predict the type of conversation based on these patterns

Table 1. *Acoustic feature descriptors: 38 low-level descriptors with regression coefficients, 21 functionals. Abbreviations: LSP: line spectral pairs, Q/A: quadratic, absolute*

Acoustic features	Functionals
PCM loudness	Position max./min.
MFCC[0-14]	arith.mean, std.deviation
log Mel Freq. Band[0-7]	skewness, kurtosis
LSP frequency[0-7]	lin. regression coeff
F0	lin.regression error Q/A
F0 envelope	quartile 1/2/3
Voicing Prob.	quartile range 2-1/3-2/3-1
Jitter local	percentile 1/99
Jitter consec. frame pairs	percentile range 99-1
Shimmer local	up-level time 75/90

with reasonable accuracy [11]. Motivated by these, Table 2 shows the conversational features used in this work to summarize the statistics on speech turn-taking, pauses and speech overlaps of the phone call.

Table 2. *Conversational features used in this work*

Conversational feats	Functionals
Call length	caller, callee, pauses and overlap
Speaker turn length	min., max and average
Overlapping speech	min., max., average
Interruption	Probability
Pauses	min. max. and average

Note that these features, unlike the acoustic features, need to be extracted for the whole recording. In total a 13-dimensional feature vector is obtained per call. The total length of the call is split into four types by considering the time only one of the two persons speaks, pauses and overlaps between both speakers. All four features are normalized to a sum of 1. Then, for each of these features we compute their minimum, maximum and average durations in the call. We also compute the probability of caller interrupting the callee, as the ratio between number of times the caller starts to speak when the callee is still speaking to the total number of times the caller starts to speak in the call.

3.3. Social Relationship Classification

The classification performed in this paper is based on the boosting framework using the BoostTexter toolkit[1] [13] with simple decision stumps as a base classifier [14]. Boosting has been successfully applied in the past for a number of machine learning tasks with highly accurate results. The core of the boosting framework calculates a weighted combination of numerous weak classifiers to form a final hypothesis that is more accurate than any of the individual classifications. At each boosting iteration, misclassified samples are given higher weights so that the classifier performing well on these samples is given more importance in the hypothesis. For the tests performed in this paper, the boosting based classifier is trained in two different scenarios, using only acoustic features and using only turn-taking features. Finally, a combined decision is also considered, as described below.

In the first scenario, where only acoustic features are used for training, the classifier predicts the type of the callee at the speaker turn level as we extract a prosodic feature vector for each turn of the caller. It is worth mentioning that the number of turns of a caller used for training is usually unbalanced between classes (partner/family). The final callee type prediction, that is, at the call level, is obtained by performing a majority voting among all the turns of the caller in the call. The number of boosting iterations is fixed empirically given a subset of data. In such tests we observed that increasing the number of iterations above 1000 did not make much difference in the performance. So while training the classifier on just acoustic features, the number of iterations is fixed to 1000. In the second scenario, the classifier directly predicts the callee type at the call level, since the turn taking features are extracted from the whole call. The number of iterations for training the classifier on conversational features was fixed to 100 empirically. To combine the outputs from the classifiers trained on acoustic and conversational features, we perform a score level averaging of the classifier outputs. Since the acoustic feature based classifier predicts the classes at turn level, we average the scores for all the turns for each class to get a single score for each class at the call level. Then, the scores obtained from the classifier trained on conversational features and acoustic features are averaged to get a final combined score for each class in the call. The label of the call is predicted by picking the class with highest score.

4. EXPERIMENTS AND RESULTS

All experiments in this paper were performed on the subset of phone calls from the CallNotes dataset as described in section 2 defining a binary classification task between the partner and family types. To evaluate the classification performance we compute the accuracy of correctly classifying a given phone call. Given that the number of calls per category is usually unbalanced both at the speaker level and globally, unless specified, we use the weighted accuracy, which takes into account the prior classification probability for each class.

In a first experiment, we explore whether it is possible to train a classifier on the data from a single caller into partner/family classes. This is done in a leave-one out classification framework. We train a boosting based classifier using all calls for a given speaker, except for one, and then try to predict the type of relationship with the callee on this unseen call. We do this for all calls and average the results. Note that when doing this we are not using any data from the test call

[1] www2.research.att.com/~astopen/download/ref/boostexter/boostexter.html

to training the classifier in order to make sure that the classifier does not learn channel or transient characteristics of the speaker from each particular call. We report in Figure 1 the results of this experiment per speaker and using the different sets of features described in Section 3.3. Next to the scores, we indicate the random classification score for each speaker, computed as the ratio of the most predominant type of relationship (i.e. the score we would get if the system always returned the most predominant class).

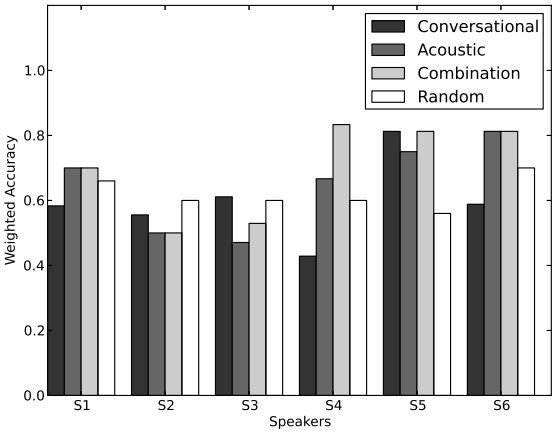

Fig. 1. *Speaker-dependent weighted accuracies of a classification task between partner and family for different features.*

We see how for most speakers we can clearly determine the social relationship with their counterpart using the proposed method. As expected, the variation of speaking style is not the same for all speakers. This is why speakers 2 and 3 achieve poor results (lower than baseline) and speakers 4, 5 and 6 can be classified much more easily. The accuracy of acoustic versus conversational features is very speaker dependent, although the combination of the two usually helps.

In a second experiment, we test wether a general model for social relationship can be built for all speakers. For this purpose, we train a general classifier to identify the type of relationship with the callee by pooling the data from all the callers. Similar to the first experiment, we follow an iterative leave-one out classification framework where a model is trained with the data from all speakers except for one, and then classification is performed on each one of the calls of this speaker. The results of this experiment are summarized in Table 3 and shown per speaker in Figure 2.

Table 3 shows that, on average, the conversational features (in spite of being of much lower dimensionality and obtained in an automatic manner) achieve better general classification results than the acoustic features. In addition, the combination of both types of features achieves the best results. Results in Figure 2 are similar from those in Figure 3, which prove that there is a common set of traits that change in all

Table 3. *Common classifier for all the speakers but using different feature set.*

System features	UA	WA
Acoustic	0.69	0.72
Conversational	0.72	0.73
Combination	0.74	0.77
Random	0.50	0.62

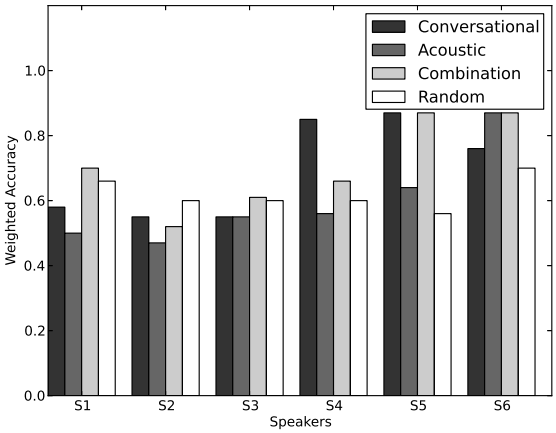

Fig. 2. *Leave-one-out speaker accuracies (speaker-independent) for different feature sets.*

speakers when addressing their partners versus their families. A strange behavior is seen in speaker s4, whose acoustic features are not discriminant when modeled alone, but become very good in the general model. This effect needs further investigation.

5. CONCLUSIONS AND FUTURE WORK

In this paper we proposed an automatic method to detect the social relationship between a speaker and the people in his social network based on the analysis of their phone conversations by using only the acoustic characteristics of the caller and the conversational characteristics extracted automatically from the calls. Every phone call is initially preprocessed automatically to extract a set of conversational features for the call and acoustic features for each of the speech segments from the caller. A boosting algorithm is then used to train a classifier by using these features. Tests are performed on a binary classification task using a real-life phone calls dataset where we show results of around 75% accuracy on classifying calls according to social relationship. As a next step we are planning to identify what particular characteristics are most important for the classification, and to collect more audio data to test the system on a multi-class classification problem.

6. REFERENCES

[1] Stephen W. Littlejohn and Karen A. Foss, *Theories of Human Communication*, 2010.

[2] Richard Y. Bourhis, Howard Giles, and Wallace E. Lambert, "Social Consequences of Accommodating One's Style of Speech: a Cross-National Investigation," *International Journal of the Sociology of Language*, vol. 1975, no. 6, pp. 55–72, 2009.

[3] Nick Campbell and Parham Mokhtari, "Voice Quality : the 4th Prosodic Dimension," in *Proc. 15th ICPhS*, 2003, pp. 203–206.

[4] Nick Campbell, "Changes in Voice Quality due to Social Conditions," in *Proc. 16th ICPhS*, 2007, number August, pp. 2093–2096.

[5] Nick Campbell, "Whom we laugh with affects how we laugh," in *Proc. of the Interdisciplinary Workshop on The Phonetics of Laughter*, 2007, number August, pp. 61–65.

[6] Anthony Stark, Izhak Shafran, and Jeffrey Kaye, "Hello , Who is Calling ?: Can Words Reveal the Social Nature of Conversations ?," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 112–119.

[7] Juan Pablo Carrascal, Rodrigo de Oliveira, and Mauro Cherubini, "A note paper on note-taking: understanding annotations of mobile phone calls," in *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. ACM, 2012, pp. 21–24.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 1459–1462, ACM.

[9] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Müller Christian, and Shrikanth Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge German Research Center for Artificial Intelligence (DFKI), Saarbr ," in *Proc. Interspeech*, 2010, number September, pp. 2794–2797.

[10] Fabio Valente and Alessandro Vinciarelli, "Language-Independent Socio-Emotional Role Recognition in the AMI Meetings Corpus," in *Proc. Interspeech*, 2011, pp. 3077–3080.

[11] Kornel Laskowski, Mari Osterdorf, and Tanja Schultz, "Modeling vocal interaction for text-independent classification of conversation type," in *8th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, Antwerpen, Belgium, 2007, pp. 194–201.

[12] Kornel Laskowski, Mari Osterdorf, and Tanja Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," in *9th ISCA/ACL SIGdial*, Columbus, USA, 2008, pp. 148–155.

[13] Robert E. Schapire and Yoram Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135–168, 2000.

[14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, pp. 2000, 1998.