A corpus for identification of speakers and their emotions

Marie Tahon, Agnes Delaborde, Claude Barras, Laurence Devillers

A corpus for identification of speakers and their emotions

Marie Tahon, Agnès Delaborde, Claude Barras, Laurence Devillers
LIMSI-CNRS
BP 133, 91 403Orsay cedex, France
E-mail: (mtahon | agdelabo | barras | devillers)@limsi.fr

Abstract

This paper deals with a new corpus, called corpus IDV for "Institut De la Vision", collected within the framework of the project ROMEO (Cap Digital French national project founded by FUI6). The aim of the project is to construct a robot assistant for dependent person (blind, elderly person). Two of the robot functionalities are speaker identification and emotion detection. In order to train our detection system, we have collected a corpus with blind and half-blind person from 23 to 79 years old in situations close to the final application of the robot assistant. This paper explains how the corpus has been collected and shows first results on speaker identification.

1. Introduction

The aim of the project ROMEO (Cap Digital French national project founded by FUI6, http://www.projetromeo.com) is to design a robotic companion (1m40) which can play different roles: a robot assistant for dependent person (blind, elderly person) and a game companion for children. The functionalities that we aim to develop are speaker identification (one speaker among N, impostor) and emotion detection in every day speech. The main challenge is to develop a strong speaker detection system with emotional speech and an emotion detection system knowing the speaker. All our systems are supposed to be real time systems.

In the final demonstration, the robot assistant will have to execute some tasks as defined in a detailed scenario. The robot is in an apartment with its owner, an elderly and blind person. During the whole day, the owner will have some visitors. The robot will have to recognize who are the different characters: his little children (two girls and a boy), the doctor, the house-keeper and an unknown person. In the scenario the robot will also have to recognize emotions. For example, Romeo would be able to detect how the owner feels when he wakes up (positive or negative) and to detect anger in the little girl's voice.

To improve our detection systems (speaker and emotion) we need different corpora, the closer to final demonstration they are, the better the results will be. We focused on blind or half-blind speakers (elderly and young person) and children voices while they interact with a robot (Delaborde et al., 2009) in order to have real-life conditions. However, emotions in real-life conditions are complex and the different factors involved in the emergence of an emotional manifestation are strongly linked together (Scherer, 2003).

In this paper, we will describe the IDV corpus which was collected with blind and half-blind person: acquisition protocol, scenarii involved. Then we explain the annotation protocol. And in section 4, we give our first results on speaker identification (identify a speaker from a set of known speakers).

2. IDV corpus

The part of the final scenario that concerns IDV corpus, we aim to demonstrate at the end of the project consists in:

- identify a speaker from a set of known speakers (children or adults),
- recognize a speaker as unknown and in this case, provide its category (children, adult, elderly) and gender (for adults only),
- and detect positive or negative emotion.

Speaker identification and emotion detection are real time tasks. For that objective, we have collected a first corpus called IDV corpus with blind and half-blind French people from 23 to 79 years old. This corpus has been collected without any robot but a Wizard-of-oZ which simulates an emotion detection system. This corpus is not fully recorded yet; further records are scheduled with the IDV. A second corpus will be collected in the context of the scenario: at the IDV (Institut de la Vision in Paris) with the robot ROMEO.

2.1 Corpus characteristics

So far, we recorded 10h48' of French emotional speech. 28 speakers (11 males and 17 females) were recorded with a lapel-microphone at 48kHz. In accordance with the Romeo Project target room, the recordings took place in an almost empty studio (apart from some basic pieces of furniture), which implies a high reverberation time. The originality of this corpus lies in the selection of speakers: for a same scientifically controlled recording protocol, we can compare both young voices (from 20 years old) to voices of older person (so far, the oldest in this corpus is 89).

2.2 Acquisition protocol

Before the recording starts, the participant is asked some profile data (sex, location, age, type of visual deficiency, occupation and marital status). An experimenter from the LIMSI interviews the volunteer following three sequences described below in 2.3.

Some parasite noise happened to be audible in the studio (guide dog walking around, people working outside, talking, moving in the corridor, …). When overlapping the speaker's speech, these parts were discarded.

2.3 Sequences description

Each recording is divided into three sequences.

The first one is an introduction to the Romeo project: we explain the participant that we need him to provide us with emotional data, so that we can improve our emotion detection system in a future robot. We take advantage of this sequence to calibrate the participant's microphone. Since there is no experimental control over the emotions that could be expressed by the participant, this part is discarded in the final corpus and will not be annotated.

In the second sequence, called “words repetition” (table 1), the experimenter asks the participant to repeat after him orders that could be given to the robot. The participant is free to choose the intonation and the expression of his or her production. This sequence gives us a sub-corpus where lexicon is determined and emotions mainly neutral.

Viens par ici! (come here!)	Mets le plat au four! (put the dish in the oven!)
Arrête-toi! (stop here!)	Descends la poubelle! (Bring down the bin!)
Stop! (stop!)	Va chercher le courrier! (Bring back the mails!)
Ecoute-moi! (listen to me!)	Va chercher à boire! (Bring back some water!)
Approche! (come near!)	Aide-moi à me lever! (help me to get up!)
Va-t-en! (go away!)	Aide-moi à marcher! (help me to walk!)
Donne! (give it!)	
Roméo, réveille-toi! (Romeo, wake up!)	
Ramasse ça! (pick it up!)	

Table 1: List of words and expressions in French

In the third sequence, called “scenarii”, the experimenter presents six scenarii (see table Scenarii) in which the participant has to pretend to be interacting with a domestic robot called Romeo. For each presented scenario, the experimenter asks the participant to act a specific emotion linked to the context of the scenario : for instance Joy, "Your children come to see you and you appreciate that, tell the robot that everything is fine for you and you don't need its help", or Stress, "You stand up from your armchair and hit your head in the window, ask Romeo to come for help", or Sadness, "You wake up and the robot comes to ask about your health. You explain it that you're depressed". The participant has to picture himself or herself in this context and to speak in a way that the emotions are easily recognizable. He or her knows that the lexicon he uses is not taken into account; the emotion has to be heard in his or her voice.

At the end of each of his or her performance, the experimenter runs a Wizard-of-Oz emotion detection tool, that tells aloud the recognized emotion. The system is presented as being under-development, and most of the times it does not correctly recognize the emotion: it

can recognize an emotion that is of the opposite valence of what the participant was supposed to express (the experimenter selects Anger when Joy has been acted); it can recognize no emotion at all (the experimenter selects Neutral when a strong Anger was expressed, or when the emotion has not been acted intensely enough); it can recognize an emotion that is close to what is expected, but too strong or too weak (Sadness instead of Disappointment). The participant is asked to act the emotion again, either until it is correctly recognized by the system, or when the experimenter feels that the participant is tired of the game.

Emotional data acquired through acting games obviously do not reflect real-life emotional expressions. However, the strategies that are being used through our Wizard-of-Oz emotion detection tool allow us to elicit emotional reaction in the participant. An example: the participant is convinced that he expressed Joy, but the system recognizes Sadness. The participant's emotional reactions are amusement, or frustration, boredom, irritation. Our corpus is then made of both acted emotions, and spontaneous reactions to controlled triggers. The distinction between acted and spontaneous expressions will be spotted in our annotations; this distinction is really important to have an estimation of how natural the corpus is (Tahon, Devillers, 2010).

We can also question the relevancy of having the participant imagine the situation, instead of having him live it in an experimental setting. We should note that for obvious ethical reasons we can not put them in a situation of emergency such as "being hurt, and ask for immediate help": we can only have them pretend it. Another obvious reason for setting this kind of limited protocol is a matter of credibility of the settings: currently, the only available prototype does not fit the target application characteristics (Nao is fifty centimeters high, and its motion is still under development).

Scenarii	**Emotions**
Medical emergency	Pain, stress
Suspicious noises	Fear, anguish, anxiety
Awaking (good mood)	Satisfaction, joy
Awaking (bad health)	Pain, irritation, anger
Awaking (bad mood)	Sadness, irritation
Visit from close relations	Joy

Table 2: Scenarii

Table 2 summarizes the 6 different scenarii and the emotions asked to the participant.

3. Corpus annotations

3.1 Emotion labels

Segmentation and annotation of the data are done with the Transcriber annotation tool[1] on the scenario sequences.

The participant utterances are split into emotional segments. These segments mark the boundary of the emotion: when a specific emotion expression starts, and when it comes to an end.

On each segment, three labels describe the emotion. The first label corresponds to the most salient perceived emotion, while the two others characterize more precisely the emotion, balance it. The table Emotion Labels presents the annotation emotional value that are used.

Other dimensions are annotated :

- Intensity: the strength of the emotion, 5 scales from *very weak* to *very strong*.
- Activation: how many different phonatory means are involved to express the emotion (voice trembling, change in loudness…), 5 scales from *very few* to *a lot*.
- Control: does the speaker contain the expression of the emotion, 5 scales from *not at all* to *completely.*
- Valence: does the speaker feel positive or negative ? *positive*, *negative*, *positive and negative*, *either positive or negative*, *valence* indeterminable.
- Audio quality: if the recorded segment quality is fine or not (microphone noise, participant speaking too close…), from *good* to *bad*.
- Spontaneous/Acted : a simple flag meant to spot if the participant was at that time acting an emotion in the context of a scenario, or reacting spontaneously to an event.

[1] http://trans.sourceforge.net/en/presentation.php

Macro-classes	Annotations values
POSITIVE	Joy
	Amusement
	Satisfaction
ANGER	Anger
	Irritation
SADNESS	Sadness
	Disappointment
FEAR	Fear
	Anxiety
	Stress
	Embarrassment
NEUTRAL	Neutral
	Positive
	Negative
	Surprise
OTHERS	Irony
	Compassion
	Interest
	Scorn
	Boredom
	Pain
	Motherese
	Excitation

Table 3: Emotion labels

3.2 IDV emotional content

As the emotional annotation of the IDV corpus is not finished yet, all results on emotion annotation are based on a set of 15 speakers.

IDV corpus is divided in two different corpora: spontaneous and acted, according to the task (as defined in part 3). The results of the emotion scores are reported in table 4.

The spontaneous corpus contains 736 instances of 0.5s to 5s. The most important emotional label is "interest" (51%). This corresponds to the agreement of the volunteer with what the interviewer asked him to do. Positive emotions (18%) are more numerous than negative emotions (6%). The volunteer has accepted to be recorded, so he is not supposed to express displeasure, he will more probably be nice with the LIMSI team. Macro-class "fear" is also quite important (10%). It corresponds to embarrassment or anxiety, playing the actor is not an easy task.

The acted corpus contains 866 instances of 0.5s to 6s. The results corresponds to what was expected: the main emotions are well represented. Positive emotion (21%, mainly "satisfaction"), negative emotion (24%, mainly "irritation"), fear (10%, mainly anxiety) and sadness (8%, "deception" and "sadness").

Label emotion	Spontaneous	Acted
joy	0.82	5.54
amusement	8.49	0.58
satisfaction	4.89	10.51
POSITIVE	**14.2**	**16.63**
anger	0.2	3.23
irritation	1.9	15.01
ANGER	**2.11**	**18.24**
sadness	1.63	3.98
deception	2.72	4.33
SADNESS	**4.35**	**8.31**
fear	0.07	4.97
anxiety	6.66	16.28
stress	1.56	2.89
embarrassment	1.22	0.06
FEAR	**9.51**	**24.19**
neutral	7.2	2.25
positive	3.94	4.16
negative	1.36	4.73
surprise	4.14	2.37
NEUTRAL	**16.64**	**13.51**
irony	0.07	0.23
boredom	2.17	1.04
interest	50.54	12.93
pain	0.07	3.87
excitation	0.07	0.46
motherese	0.27	0.46
OTHERS	**53.19**	**19.11**

Table 4: Emotion scores (%) for both spontaneous and acted IDV corpora

4. IDV first results

In this section, speaker identification scores are presented. All the results presented here were obtained with the same method based on GMM (Gaussian Mixture Models) speaker models (Reynolds et al., 2000). First we have studied the different parameters of the GMM model, then the evolution of scores in function of the sex and the age of speakers.

4.1 Global speaker identification scores

This section aims at choosing the experimental setup for studying the influence of the age, gender and emotional expression. Experiments are performed with the "repeating words" sequence of the corpus. It contains 458 audio segments of varied duration. 26-dimensional acoustic features (13 MFCC and their first-order

temporal derivatives) are extracted from the signal every 10ms using a 30ms analysis window. For each speaker, a training set is constructed by the concatenation of segments up to a requested duration Ntrain; a Gaussian mixture model (GMM) with diagonal covariance matrices is then trained on this data through maximum likelihood estimation with 5 EM iterations. The remaining segments, truncated to a Ntest duration, are used for the tests. For a given duration, the number of available segments is limited by the number of segments already used for training and the minimal test duration necessary (the higher duration is, the less audio files there are). For each test segment, the most likely speaker is selected according to the likelihood of the speaker models.

In order to optimize the number of files of train and test, we have chosen the following set of parameters:

- test duration: 1s (225 files),
- train duration: 10s (179 files),
- speaker model: mixture of 6 Gaussians.

The error rate is 34.7% (+/- 6.5%) when recognizing one speaker among 28.

This extremely short test segment duration is due to constraints on segment counts in the database, and improvement of the performance as a function of the segment length will be studied later in the course of the project.

4.2 Age influence

In this part, we show that speaker identification is easier on elderly person voices than on young voices. Two sub-corpora from IDV corpus composed of the 8 older volunteers (4 male, 4 female, from 52 to 79 years old), respectively the 8 younger volunteers (4 male, 4 female, from 23 to 46 years old) are studied separately. Of course, the number of segments is quite low, which may be a bias of the experiment.

The results are referred in the table 5, error rate, number of segments for test and trust interval (binomial distribution test).

	Old person	**Young person**
Error rate	17.00%	38.00%
Number of segment	66	63
Trust interval	9.18%	12.24%

Table 5: Speaker identification, age influence: error rate, number of segment and trust interval

As a result speaker identification (one speaker among N) is better with elderly person voices. Our hypothesis is that voice qualities are much more different with elderly person voices than with young voices. In figure 1, we have plotted the MFCC2 gaussian model for the first four older person (blue) and for the first four younger person (red). As the red curves are quite the same, the blue one are more separated one from another.

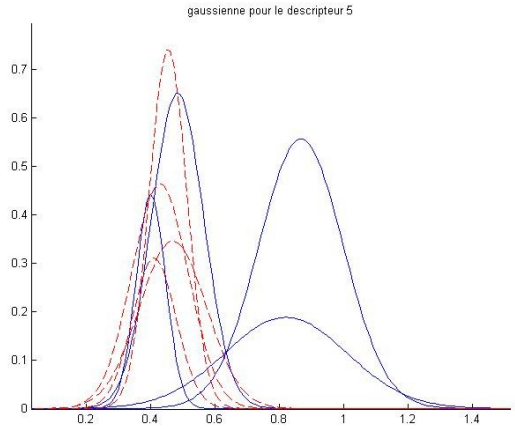

Figure 1: Distribution of the 4th MFCC coefficient according to a gaussian model for old (blue, plain) and young speaker (red, dashed)

4.3 Sex influence

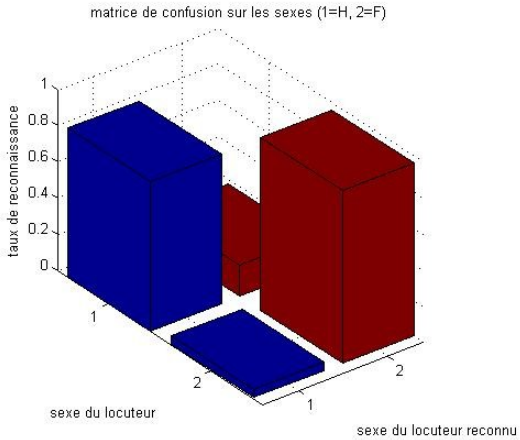

Figure 2 : Confusion matrix between male (1) and female (2)

Based on the whole IDV corpus, we compute the confusion matrix sorted by sex without taking into account the age of the speakers anymore.

A female voice is recognized as well at 96%, a male voice is recognized as well at 82%. Female voices have

better identification scores.

4.4 Emotional speech influence

The results below are based on the corpus "repeating words" which contains 28 speakers. The results presented in this part are based on both sequences "repeating words" and "scenario", with the 15 speakers corresponding to the emotional annotation of the sequence "scenario". Table 5 below shows the error rate for speaker identification (1 among 15) across the 3 corpora: "repeating words", "scenario spontaneous" and "scenario acted". The parameters we have chosen for the gaussian model are the followings: 5 gaussians, train duration: 10s, test duration: 1s.

		TEST		
		"Words"	**"Spontaneous"**	**"Acted"**
TRAIN	**"Words"**	28.60%	78.60%	88.00%
	"Spontaneous"	X	45.10%	60.20%
	"Acted"	X	X	56.30%

Table 5: Error rates for speaker identification across the three corpora

Identification scores are better with the "words" corpus (lexically controlled) than with the "acted" corpus. The "spontaneous" corpus gives intermediate results. The scores are always better when the train and the test are made on the same corpus.

Speaker models were tested directly in mismatch conditions without any specific adaptation. The very high error rates observed are of course due to the very short train and test durations constraints in our experiments, but also highlight the necessity of an adaptation of the speaker models to the emotional context which will be explored during the ROMEO project.

5. Conclusion

This corpus IDV is interesting for many reasons. First, as it presents a sequence of words, lexically determined by the protocol and quite neutral, and a sequence of emotional speech, with the same speakers, recorded in the same audio conditions, it allows us to compare scores for speaker identification between neutral speech and emotional speech. Secondly, the corpus collection has been made with blind and half-blind volunteers from 23 to 79 years old. Thus we can compare scores across speaker age. Moreover we have the opportunity to work with elderly person who often have specific voice qualities.

6. References

Delaborde A., Tahon M., Barras C., Devillers L. (2009). A Wizard-of-Oz game for collecting emotional audio data in a children-robot interaction. AFINE 2009.

Tahon M. and Devillers L. (2010). Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Proceedings of the Fifth International Conference on Speech Prosody*, 2010.

Devillers L., Abrilian S., Martin J-C., (2005) Representing Real-life Emotions in Audiovisual Data with Non Basic Emotional Patterns and Context Features, ACII 2005.

Devillers, L. Vidrascu, L. Lamel, L. (2005). Emotion detection in real-life spoken dialogs recorded in call center, *Neural Networks*, Special Issue on "Emotion and Brain", ELSEVIER, Vol. 18, No. 4, pp. 407-422, 2005

Reynolds D., Quatieri T., and Dunn R. (2000) Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

Scherer K. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication*, vol. 40, pp227-256, 2003;