Emotional speaker identification from the vocal characteristics

Malika Akak, Halim Sayoud

الملخّـص

هذا المقال يقدّم التّعرف على الحالة النّفسيّة للمتكلم، باعتماد قاعدة معطيات. اقتُرح أحد عشر عاملا متعلّقا بالتنغيم، لاستخراج مؤشّرات لإشارة الكلام، كما اعتمدت طرق إحصائية لتقليل حجم المعطيات المستعملة.

الكلمات المفاتيح : إشارة الكلام، التّنغيم، التّعرف الآلي على الكلام، الحالة النّفسيّة للمتكلم.

Emotional speaker identification from the vocal characteristics

Malika Akak[1], Halim Sayoud[2]

[1]Electronic Department Saad Dahleb University, Blida, Algeria
[2]LCPTS Department, U.S.T.H.B, Algiers, Algeria.

akak.malika@yahoo.fr, halim.sayoud@gmail.com

Abstract

This paper presents the identification of emotional state of the speaker from a variety of database given. Eleven prosodic parameters have been proposed for features extraction and statistical techniques are used to minimize the amount of data to be handled.

Keywords: speech emotion, speech recognition, speech signal, prosody.

1. Introduction

The speech is extremely rich and complex. It is a mean of communication that translates not only the linguistic information but also the information about the identity, the personality and the emotional state of the speaker. Recent development has made it possible to use this in the security system. In emotion identification, the task is to use a speech sample to select the emotional state of the person that produced the speech from among a database of speakers. This technique makes it possible to use the speakers' voice in many domains for instance to detect the lies in criminal investigations and the problems of understanding the dialog between speaker and the man-machine' system which allows to take care of the customer relationship and to answer more quickly, more economically the requests of the customers to avoid their anger [1, 8, 11-12].

2. Generals principles identification

Emotion identification methods can be divided into text-independent and text-dependent methods. In a text-independent system, emotional models capture speech characteristics of somebody which show up irrespective of what one is saying. In a text-dependent system, on the other hand, the identification of the speaker's emotion is according to his or her speaking of special text, one or more specific phrases, like passwords, card numbers, PIN codes, etc. Each method has its own advantages and disadvantages and may require different treatments and techniques [1, 7-8].

2.1. Speech features extraction

The purpose of this section is to convert the speech waveform to some type of parametric presentations. The speech signal is considered as quasi-stationary, when examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. For long periods of time (on the order of 0.2s or more) the signal characteristics change to reflect the different speech sounds being spoken. Therefore, short-time spectral

analysis is the most common way to characterize the speech signal [10].

A wide range of possibilities exist for parametrically representations of the speech signal for the emotion identification task, such as pitch, energy, pauses, and others. Pitch, energy and pauses are perhaps the best known and most popular. These features have been used in this paper with their derived, (maximal, minimal, jitter, average etc.). Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Our eleven parameters are more affected by the emotional state of the speaker. Every change of the emotional state modifies the fundamental frequency vector, energy and pauses. For instance an anger speaker speaks quickly with high volume. He does fewer pauses with short durations. An agitated speaker speaks slowly and does more pauses with low level of energy. A person who has a strong character (or who is confident) trusts on her or himself and speaks normally with a normal state of pitch, energy and pauses. A sad man speaks very slowly with less energy and does many very long pauses. For this reason, it is clear that the emotional state of the speaker depends on all parameters which are extracted from the fundamental frequency or from the energy and pauses characteristics. We vary our fundamental frequency as to express our feelings and to attract the attention of the listener to important aspect in our speech message. The paragraph pounced with a constant and uniform pitch appears fairly natural [2,5,8,9, 11,12].

For the aim of the feature extraction section is to obtain a new voice representation which is more compact, less redundant, and more suitable for statistical model. The purpose of our search is to prove that the change of the emotional state affects our parameters, with variable proportions and to measure the pertinence rate of every parameter and emotion.

2.2. Proposed statistical prosodic parameters

This part briefly describes our parameters which convert the speech signal to eleven prosodic parameters. Now we can say that the pitch, energy and pauses are important. Their variations are influenced by the emotional state of the speakers that's why we have chosen the prosodic parameters in function of the pitch, energy and pauses and we have thought to use their derived.

2.2.1. The first parameter

Is the speed of the speech signal which represents the difference between both maximal and minimal values of F_0. It is symbolized and calculated by formula:

$$P_1 = \max(F_0) - \min(F_0) \quad (1)$$

Such as $\max(F_0)$ and $\min(F_0)$ are respectively the maximal and minimal values of F_0.

2.2.2. The second parameter

It represents the maximal speed of the speech signal? This parameter is symbolized and calculated according to the formula (2).

$$P_2 = \max(F_0(j) - F_0(i)) \quad (2)$$

$F_0(j) - F_0(i)$ designs the successive variation between two neighbors values of F_0.

2.2.3. The third parameter

It represents the minimal speed value of the speech signal, symbolized and calculated according to the formula (3):

$$P_3 = \min(F_0(j) - F_0(i)) \quad (3)$$

2.2.4. The fourth parameter

Is the global energy of the speech signal which is symbolized and computed by the following formula?

$$P_4 = \log \sum_{i=1}^{N} |S(i)| \quad (4)$$

With N is the size of the analysis window which is equal at10 ms [8-9].

2.2.5. The fifth parameter

It represents the maximal energy of the speech signal symbolized and calculated by the formula (5)

$$P_5 = \max \left[20 \frac{\log \sum_{i=1}^{N} |S(i)|}{N} \right] \quad (5)$$

With N is the size of the analysis window which is equal to 10ms in our case [8-9].

2.2.6. The sixth parameter

It indicates the signal variation in decibels which is symbolized and calculated by the formula:

$$P_6 = 20 \log \left[\max \sum_{i=1}^{N} |S(i)| - \min \sum_{i=1}^{N} |S(i)| \right] \quad (6)$$

2.2.7. The seven parameter

It indicates the quantitative description of the voice quality called by Jitter which is symbolized and defined by the formula (7):

$$P_7 = \frac{100 N_0}{(N_0 - 1) \sum_{i=1}^{N} |S(i)|} \sum_{i=1}^{N-1} |T_i - T_{i+1}| \quad (7)$$

Where $T_i = 1/F_i$ and F_i represents the value of the ith frame of the fundamental frequency. N0 is the total number of the voiced frames in the speech signal [13].

Before beginning of the other parameters concerning the silent pauses, it was necessary to eliminate the silent segment at the beginning and the end of the signal. Then to segment the signal in two parts: silent and speech.

2.2.8. The eighth parameter

It represents the number of the pauses in the speech signal. This parameter is symbolized by P_8.

2.2.9. The ninth parameter

It represents the total durations of the pauses in the speech signal. It is symbolized by P_9.

2.2.10. The tenth parameter

It represents the maximal of the pauses in each speech signal, symbolized by P_{10}.

2.2.11. The eleventh parameter

It represents the average durations of the pauses in the speech signal. It is symbolized by P_{11}.

3. Principle aims

It consists of four separate phases. The first phase extracts all our parameters and computes the margins between minimal and maximal values for each parameter, where the four desired emotions fixed in

advance (anger, agitation, confident and strong character). The second phase selects threshold separately. The thresholds (if they exist) must assure the more possible separation between emotional and the no emotional speech signal for each parameter and emotion. The next phase computes the identification rate which symbolizes the true identification of the emotional speech signal rate, the false acceptance and the false rejection rates are used in the computing of the equal error rate as the half of the their addition . The last phase takes the minimal value of equal error rate which corresponds on the maximal value of the emotional identification rate (for each emotion and parameter).

In the identification phase our model detects the emotion represented by a sequence of feature vectors. The detected emotion is compared to the naïve emotion. For each speech signal four kind of distortions measure are computed using the threshold for each parameter (if they exist). The lowest distortion which minimizes the equal error rate is chosen. This value corresponds to a maximal of the emotional identification rate. The parameters correspond to theses thresholds which have been chosen are the pertinent parameters for the identification. Every emotion has a particular threshold. General's blocs of the identification are illustrated by figure 1.

3.1. Emotional identification rate

Is expressed and symbolized by the following formula in equation (8):

$$ID(\%) = \frac{N_1}{N_T} \times 100 \tag{8}$$

N_1 indicates the number of the true identification and N_T is the total number of the speech signal.

3.2. The false acceptance and the false rejection

In biometric domain, the false acceptance to the no authorized users (in identifying and in verifying), is considered as the most grave biometric security system error (type II) because the users have been identified by the security system instead of to avoid them. A false rejection to the authorized users by the security system instead of to allow them, is considered as error type I.

The false acceptance and rejection rates were computed using the probability measurements. [3-4].

False alarm rate is indicated by the following formula in equation (9):

$$FA(\%) = 100 \frac{N_2}{N_3} \tag{9}$$

N_2 symbolizes the number of the false alarm and N_3 symbolizes the number of the emotional speech signal.

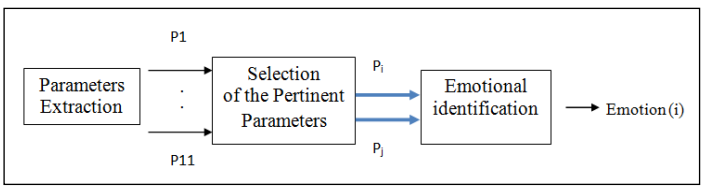

Figure 1: The stages of the identification

Missed detections rate is indicated by the equation (10):

$$MD(\%) = 100\frac{N_4}{N_T} \quad (10)$$

N_4 indicates the number of the no detected emotional speech signal and N_T is the total number of the emotional speech signal.

3.3. Equal error rate (EER)

In function to the threshold values the false acceptance and rejection are computed. The value of equal error rate corresponds to the equality of the false acceptance and the rejection rates. In practical the equal error rate is computed by the formula (11) [3-4, 7]:

$$EER\ (\%) = \frac{MD + FA}{2} \quad (11)$$

3.4. Estimation of the equal error rate

The choice of the threshold value is crucial because the function of the identification system will be influenced by this choice. The more the value of the threshold is low, the more the false acceptance rate will be important. The more the value of the threshold is important, the more the false rejection rate will be important. In practical the value of equal error rate estimated by the half of the addition of the false acceptance and rejection rates [3-7].

4. Database

Our Data Base is a set of the German emotional speech signal which is called ORATOR. Their verbal contents are constant texts which are a variety of 8 different sentences. The duration of every sentence varies from 1 to 11 seconds. The lexical information does not influence the emotional identification because the same sentences were used with the unknown language of expression. One part of ORATOR is detailed in this experiment. It concerns the male speakers.

5. Results

This part exposes all results of our experience. Tables 1, 2 and 3 show the corresponding thresholds.

	Agitation	Anger
P_1	149	150
P_2	33	/
P_3	33	/
P_4	145	145
P_5	89	90
P_6	218	218.5
P_7	3	3
P_8	2	1
P_9	3	2
P_{10}	2	2
P_{11}	1.5	1

Table 1. Thresholds for agitation and anger.

	Confidence	Strong charater
P_1	140	140
P_2	46	/
P_3	/	/
P_4	140	144

Table 2. Thresholds for confidence and strong (Part 1).

	Confence	Strong character
P_5	88	89
P_6	216.8	217
P_7	3	3
P_8	3	2
P_9	6	5
P_{10}	4	3
P_{11}	2	2

Table 3. Threshold for confidence and strong (Part 2)

Table 4 illustrates the variation margins of the parameters. The equal error and the maximal of the emotional identification rates for each parameter are showed in the flowing tables 5 and 6. The first one is for agitation and anger emotions. The second and the third one are for the rest.

Parameter	[Minimal - Maximal]
P_1	[27,37 - 295,21]
P_2	[7,66 - 161,44]
P_3	[9,21 - 121,53]
P_4	[105,02 - 156,55]
P_5	[70,71 - 99,13]
P_6	[216,17 - 221,80]
P_7	[1,94 - 4,73]
P_8	[0 - 11]
P_9	[0 - 40]
P_{10}	[0 - 17]
P_{11}	[0 - 7,5]

Table 4. Variations of Margins

%	Agitation		Anger	
	EER	MAX(ID)	EER	MAX(ID)
P_1	30,76	81,32	28	90,10
P_2	44,86	83,51	/	/
P_3	39,09	85,71	/	/
P_4	24,35	86,81	**20,10**	**92,30**
P_5	24,99	86,81	24,54	92,30
P_6	25,63	89,01	24,54	90,10
P_7	57,04	81,11	61,37	84,44
P_8	25.63	85,71	**19.72**	**91.20**
P_9	25.63	85.71	23.94	91.2
P_{10}	25.63	85.71	28.16	91.20
P_{11}	28.84	85.71	**19.72**	**91.20**

Table 5. Agitation and Anger identification errors.

%	Confidence		Strong character	
	EER	MAX(ID)	EER	MAX(ID)
P_1	28.91	72,52	34,99	70,32
P_2	43,33	65,93	/	/
P_3	/	/	/	/
P_4	22,92	79,12	**19,99**	**83,51**
P_5	42,11	78,02	30,76	73,62
P_6	29,33	70,32	32,3	74,72
P_7	54,51	61,11	58,45	52,22
P_8	33.45	67.03	31.53	80.67
P_9	38.53	65.93	34.99	82.41
P_{10}	40.78	62.63	34.22	82.41
P_{11}	51.83	62.63	33.84	82.41

Table 6. Confidence and Strong errors.

6. Remarks and interpretations

Let us remind that the emotional identification rate expresses the rate of the well recognized emotion.

We consider only the equal error rate because it is in function of the false alarm and the Missed detections rates. It means that the equal error rate depends on their values. We call it by error.

We have noticed that:

- The thresholds differ in most of the cases from parameter to another (tables 1, 2 and 3);
- The variations of margins vary from parameter to another and for all parameters the variations of margins are more remarkable with the exception of 4 parameters which are the eighth, ninth, tenth and eleventh parameters (table 4).

The obtained results (tables 5 and 6.) show that the maximal of the emotional identification rate:

- exceeds 81 % for agitation with all the parameters;
- exceeds 90 % for anger with some parameters (the first, fourth, fifth, sixth, eighth, ninth, tenth and eleventh parameters and it is equal at 84,44 % for the same emotion with the seventh parameter;
- varies between 61,11 % and 79,12 % for confidence with all parameters only with The third parameter.
- varies between 80,67 % and 83 % for strong character with some parameters (the fourth, eighth, ninth, tenth and eleventh parameters).

The results (tables 5 and 6.) prove that:

- The second and third parameters bring nothing for the identification of anger and strong character and the third parameter brings also nothing for confidence.

Seen the values of the maximal of the emotional identification and the equal error rates (tables 5 and 6) we have concluded that:

- The first parameter is incapable to recognize all emotions according to the unacceptable values of the error rate;
- The second parameter brings nothing to identify the anger and strong character;
- The second parameter fails to recognize agitation and confidence because the values of their equal error rates are considerable;
- The third parameter brings nothing to identify the anger, confidence and strong character. This parameter is unable to detect agitation because the equal error rate is considerable (39, 19 %);
- The fourth parameter is pertinent in discrimination of anger and strong character and it is no pertinent with agitation and confidence because the values of their equal error rates are unacceptable;
- The fifth parameter is incompetent to identify all emotions because the equal error rates are undesirable;
- The sixth parameter is ineffectual to discriminate all emotions because the equal error rate are considerable;
- The seventh parameter presents a total failure with all emotions according to the values of the equal error rates;
- The eighth parameter is capable of identify anger and it is Incapable to distinguish the rest because the values of equal error rates are unacceptable;
- The ninth parameter identifies badly the agitation, anger and strong character. This parameter can't identify confidence;

- The tenth parameter expresses its failure in discrimination of confidence also it is incapable to recognize agitation, anger and strong character;
- The eleventh parameter is able to detect anger and it is unable to detect the rest of emotions.

7. Conclusion

Using emotion identification system, we can identify the emotion of the person who is speaking. In the identification stage, the distortion measures are based on the minimizing of the errors the more possible.

Firstly, the obtained results showed the effectiveness of some parameters (the fourth, eighth, eleventh parameters) with some emotions in the task of emotion detection. We have noticed too that the agitation and anger recognition rates exceed the 91 % for the same parameters that we have just mentioned above. For the fourth parameter the strong character recognition rate equal the 83.51 % (tables 5 and 6). We have thought that the prosodic parameters change according to the emotional state, age and the health of the speaker; and differ from speaker to another according to the difference between the human, the emotional state, the health and the emotional expressions which vary widely from speaker to another. We have concluded that for all parameters the variations of margins are more remarkable form the first to fifth parameters and for the rest the variations of margins are negligible. Finally, we hope that this work will contribute to the speech recognition development in general, and to the emotional detection in particular.

8. References

[1] Triki, A., " Développement d'un Système de Reconnaissance Robuste de la Parole", Mémoire de Magister, Institut d'électronique, Centre Universitaire Cheikh el–Arbi Tebessi Tebessa, 2007.

[2] Huang, X., Alex, A. and Hsiao-woien, H., " Spoken language processing: a guide to Theory, Algorithm, and System Development" United states of America: Prentice Hall PTR, 2001.

[3] http://www.griaulebiometrics.com/en-us/book/understanding- biometrics/evaluation/accuracy/matching/interest/equal

[4] http://w.webppedia.com/TERME/ equal errorrate.html

[5] Boite, R., Bourlard, H., Dutoit, T., Hancq J. and Leich, H., " Traitement de la parole ", 2000. http://books.google.fr/books.

[6] http://www.gloccal.fr/verification-et-identification-du- locuteur.html.
Dorizzi, B., " Techniques et Usages Biométriques », GET/INT Evry2204 http://-paristic.fr/TUTIRIAL/tutorial_BIO PARISTIC_05.pdf

[7] Attabi, Y., " Reconnaissance automatique des émotions à partir du signal acoustique ", Montréal, thèse de Doctorat. Ecole de technologie supérieure. Université de Québec, 2008.

[8] Audibert, N., " Prosodie de la parole expressive : dimensionnalité d'énoncés méthodologiquement controlés authentiques et actés ", thèse de Doctorat. Institut polytechnique de Grenoble, 2008.

[9] Calliope, " La parole et son traitement automatique » Dunod, 1997.

[10] Seppânen, T., Eero, V. and Juhani, T. "Prosody-based classification of émotions in spoken Finnish". In EUROSPEECH. Geneva, Switzerland, 2003.

[11] Chung, S., " L'expression et la perception de l'émotion extraite de la parole spontanée ", thèse de Doctorat, Université Paris III, Sorbonne Nouvelle, Institut de Linguistique et Phonétique Générales et Appliquées, 2000.

[12] Huang, A., Hsiao-woien, A. and Hsiao-woien, H., "Spoken language processing: a guide to Theory, Algorithm, and System Development", United states of America: Prentice Hall PTR, 2001.

[13] Cambridge Advanced Learner's Dictionary software.