INTERFACE DATABASES: DESIGN AND COLLECTION OF A MULTILINGUAL EMOTIONAL SPEECH DATABASE

Vladimir Hozjan*, Zdravko Kacic*, Asuncion Moreno, Antonio Bonafonte**, Albino Nogueiras****

*Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova17, SI - 2000 Maribor, Slovenia
(vladimir.hozjan@uni-mb.si, kacic@uni-mb.si)
** TALP Research Center Departament de Teoria del Senyal, Comunicacions Universitat Politcnica de Catalunya
Campus Nord, Edifici D5, 08034 Barcelona, SPAIN
(asuncion@gps.tsc.upc.es, albino@gps.tsc.upc.es, antonio@gps.tsc.upc.es)

Abstract

As a part of the IST project Interface ("Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented environments"), an emotional speech database for Slovenian, English, Spanish, and French language has been recorded. The database is designed for general study of emotional speech as well as analysis of emotion characteristics for speech synthesis and for automatic emotion classification purposes. Six emotions have been defined: anger, sadness, joy, fear, disgust and surprise. The neutral styles were also recorded. One male speaker and one female speaker have been recorded, except for English language where two mail and one female speaker have been recorded. All the speakers are actors. The corpuses consist of 175-190 sentences for each language. For Spanish and Slovenian databases subjective evaluation tests have been made. The recorded Interface emotional speech database represents a good basis for emotional speech analysis and is also useful in synthesis of emotional speech.

1. Introduction

Kramer (1963) reviewed a number of studies, which have demonstrated that various aspects of speaker's physical and emotional state, including age, sex, appearance, intelligence, and personality can be identified by voice alone.

The known idea in emotion research is that certain emotions are primary, and others are secondary. Secondary emotions can be explained as combination of primary emotions. This idea can be traced back to Descartes (Anscombe and Geach, 1970). There is no definitive list of basic emotions. There is, however, a general agreement on the list of basic emotions: anger, disgust, fear, happiness, sadness, and surprise (Cornelius, 1996).

In the case of speech synthesis the aim is to modify the output of speech synthesis system to produce speech, which sounds naturally emotional. This is essentially the same goal as that of actors (Johnstone, 1996). Collecting recordings of natural speech that contain natural expressed emotion is very complicated and expensive.

As a part of the IST project Interface ("Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented Environments"), an emotional speech database for Slovenian, English, Spanish, and French language has been recorded. The database is designed for general study of emotional speech as well as analysis of emotion characteristics for speech synthesis and for automatic emotion classification purposes. The partners have established common specifications which have been followed in all the languages.

2. Structure of the Interface database

Six emotions have been defined following the MPEG-4 standard for video analysis. These six emotions (see table 1) are also very often used in analysis and synthesis

Language	English	French	Spanish	Slovenian
6 MPEG4 styles + codes	A = anger T = sadness J = joy F = fear D = disgust S = surprise	A = anger T = sadness J = joy F = fear D = disgust S = surprise	A = anger T = sadness J = joy F = fear D = disgust S = surprise	A = anger T = sadness J = joy F = fear D = disgust S = surprise
Neutral variations + codes	L = neutral/soft/ slow H = neutral / loud / fast	N = neutral/normal L = neutral/soft/slow H = neutral / loud / fast	N = neutral/normal L = neutral/soft H = neutral/loud W= neutral/slow Z= neutral/fast	L = neutral/soft/ slow H = neutral / loud / fast

Table 1: The list of emotions and neutral speaking styles and their abbreviation for each language in the database.

\<database>\<spk>\<session>	
<database>	Defined as: <dbName><#><language code> i.e. INTER1FR Where: <dbName> is INTER <#> is 1 for Synthesis database <LL> is the ISO 2-letters code FR for French, EN for English, SI for Slovenian, SP for Spanish
<spk >	Defined as: <n> where <n> is M(male) or F(female) speaker.
<session >	Defined as: SESS<s><00n> Where : <s> is a character standing for the emotional style performed by the professional actor. See table 3 for details. <00n> gives the session number (n = 1, 2).

Table 2: Definition of directory structure for all languages.

LL S n X yyy. L sf	
LL	Two letter language code (fr = French, sp = Spanish, en = English, si = Slovene)
S	Emotional Style (A = anger, T = sadness, J = joy, F = fear, D = disgust, S = surprise, N = neutral/normal, L = neutral/soft/slow, H = neutral/loud/fast, $L^{(1)}$ = neutral/soft, $H^{(1)}$ = neutral/loud, W = neutral/slow, Z = neutral/fast) $^{(1)}$: for Spanish database only
n	Recording session (1 or 2)
X	Speaker gender (M or F)
yyy	Item identifier (from 001 to 175)
L	Sample Format : Linear (L)
Sf	Sampling frequency (16 = 16kHz)

Table 3: Definition of file nomenclature of Interface databases for all languages.

of emotional speech.

The neutral styles were also defined as a reference to emotional speech. Additional variations on neutral style have been proposed: slow-soft and fast-loud neutral styles for the French, Slovenian and English databases, and slow, soft, loud, and fast neutral styles for the Spanish one.

For each language, one male speaker and one female speaker have been recorded except for English language where two male and one female speaker were recorded. All speakers were actors.

2.1. Directory structure

The definition of the directory structure is the same for all languages and is presented in table 2.

2.2. File nomenclature

File names follow the ISO 9660 file name conventions (8 plus 3 characters) according to the main CD ROM standard. The template shown in the table 3 is used.

Sentence type	**Item identifier**			
	English	**Slovenian**	**Spanish**	**French**
Digits and numbers	1 to 15	1 to 15	151 to 160	1 to 25
Isolated words	16 to 35	16 to 35	161 to 184	26 to 36
Short sentences	36 to 55	36 to 55		
Middle long sentences	56 to 115	56 to 115		
Long sentences	116 to 135	116 to 135		
Paragraph text	136 to 186	136 to 190	135 to 150	
Affirmative sentences			1 to 100	37 to 155
Interrogative sentences			101 to 134	156 to 175

Table 4: The item identifiers of the corpus for English, Slovenian, Spanish, and French database.

English		Slovenian		Spanish		French	
SAMPA	frequency (%)	SAMPA	frequency (%)	SAMPA	frequency (%)	SAMPA	frequency (%)
A:	2,11	@	0,80	g	0,15	a	7,45
E	3,30	E	6,91	tS	0,25	E	5,31
{	11,18	O	6,76	J	0,30	a~	3,08
OI	1,41	S	1,17	L	0,38	o	0,99
@U	0,79	W	1,20	b	0,39	b	1,09
I@	1,84	Z	0,71	jj	0,45	S	0,56
b	1,93	a	6,30	N	0,50	d	4,65
C	0,49	a:	3,70	x	0,71	@	0,58
d	4,12	b	1,64	rr	0,78	e	5,57
D	2,75	d	3,11	f	0,92	2	5,07
e@	2,75	dZ	0,19	G	1,06	f	1,62
e	2,14	dz	0,09	d	1,07	g	0,58
eI	1,87	e:	3,13	z	1,28	i	4,77
f	1,69	f	0,48	w	1,46	e~	1,02
g	1,03	g	1,69	T	1,89	Z	1,36
h	1,47	h	1,34	B	2,19	k	3,95
i:	4,69	i	6,44	p	2,49	l	6,52
I	3,39	i:	2,96	j	2,63	m	3,46
dZ	0,75	j	4,16	u	2,75	n	2,92
k	3,57	k	3,45	m	3,44	O	2,53
l	3,84	ks	0,01	k	3,65	9	0,52
m	2,92	l	3,88	D	4,19	o~	1,68
n	7,24	l'	0,05	i	4,27	u	2,00
N	1,10	lj	0,14	t	4,65	p	3,43
O	1,35	m	3,87	l	4,93	R	8,22
o	0,08	n	6,74	r	5,35	s	5,53
p	2,22	n'	0,05	s	5,94	t	5,55
r	4,54	nj	0,11	n	6,08	H	2,27
s	5,30	o:	2,19	o	9,40	H + i	0,44
S	0,69	p	3,34	e	12,84	g~	0,27
t	7,05	r	4,77	a	13,60	v	2,39
T	0,32	s	4,87			w	0,74
u	0,38	t	4,66			j	2,14
aU	1,46	tS	1,48			z	1,73
v	2,00	ts	0,83				
w	2,13	u	1,28				
U@	0,76	u:	0,52				
z	3,29	v	2,41				
Z	0,04	w	0,77				
		z	1,82				

Table 5: The frequencies of phonemes in the corpuses for each language.

3. Corpuses

The corpuses consist of 175-190 sentences for each language. They include isolated words, sentences, and a text passage. Sentences have been chosen with different length (short, medium, and long) including affirmative and interrogative forms. The table 4 show the definitions of the corpuses for each language.

Short sentences are composed from five to eight words, middle long sentences are composed from nine to thirteen words and long sentences are composed from fourteen to eighteen words.

The sentences of the corpus were selected from large collection of text that should contain emotionally neutral context. The sentences were selected in a way to achieve phonetically balanced corpuses. Table 5 shows the frequencies of phonemes for each language.

4. Recording condition

The recordings have been performed in silent rooms using a high quality condenser microphones. The speakers have read the sentences either from prompt sheet or from the display of the PC, which has been placed on the other side of a glass window. Except for the French database, two channels were recorded simultaneously, one for the microphone and the other for the laryngograph signal.

Recordings of the English and the Slovenian database have been made using a condenser microphone, AKG 3000B and Portable Laryngograph from Laryngograph Ltd. Recordings of Spanish database have been made

	S	J	A	F	D	T	N	
S	89	20	7	0	6	2	4	**128**
J	0	115	7	0	2	2	2	**128**
A	2	14	85	2	5	5	15	**128**
F	4	1	1	103	5	13	1	**128**
D	2	1	2	5	106	3	9	**128**
T	1	3	1	16	3	101	3	**128**
N	0	2	2	1	4	1	118	**128**
	98	**156**	**105**	**127**	**131**	**127**	**152**	896

Table 6: Subjective evaluation test for Spanish database. Values in the columns represent the number of recognised sentences with the assigned emotions in the first choice for sentences belonging to the emotion of each row. A denotes anger, D disgust, F fear, J joy, S surprise, T sadness, and N neutral. In the last row the values represent the number of recognised sentences for each emotion

	A	D	J	F	S	T	N	H	
A	58	5	1	2	6	0	0	0	72
D	10	29	0	17	7	4	1	4	72
F	3	0	32	11	10	9	2	5	72
J	4	1	2	50	5	2	3	5	72
S	1	2	9	24	25	3	3	5	72
T	0	2	4	0	2	49	14	1	72
N	0	4	0	0	2	9	51	6	72
H	9	2	3	7	3	2	4	42	72
	85	45	51	111	60	78	78	68	576

Table 7: Subjective evaluation test for Slovenian database. Values in the columns represent the number of recognised sentences with the assigned emotions in the first choice for sentences belonging to the emotion of each row. A denotes anger, D disgust, F fear, J joy, S surprise, T sadness, N neutral slow, and H neutral fast. In the last row the values represent the number of recognised sentences for each emotion.

using a condenser microphone AKG 320, and of French database using AKG C391B.

5. Speech material

Each speaker recorded two sessions that were 2 weeks apart. For each session, the speaker read the whole corpus in all the emotional and neutral styles. The duration of a session for one speaker, reading the script in six emotional styles and neutral styles, was about four hours.

English Interface database contains 8928 sentences, Slovenian 6080 sentences, French 5600, and Spanish contain 5520 sentences. Different number of recorded sentences is a consequence of different number of neutral speaking styles and speakers.

6. Subjective evaluation test

For Spanish and Slovenian databases subjective evaluation tests have been made. Spanish subjective evaluation test included 16 non professional listeners (engineering students from UPC). 56 utterances were played (seven per emotion including long and short ones). Slovenian subjective evaluation test was performed by 11 non professional listeners (employees and students from UMB). 64 utterances were played (eight per emotion including long and short ones). Each listener decided which emotion corresponds to each utterance and the intensity of the perceived emotion in a one to five scale. A second choice may be marked if the first one is not clear. Test listeners never heard recorded speakers before test. Recordings of female and male speaker were played alternated, so that listeners didn't had immediate reference to previous emotion in speech.

Results of the Spanish subjective evaluation test (see table 6) showed more than 80% accurateness for the first choice. 90% accuracy if second choice was also considered. Each utterance is correctly recognised by, at least, half of the listeners. All errors affect short utterances, all the sentences and paragraphs were correctly recognised by all the listeners in first choice. Results of the Slovenian subjective evaluation (see table 7) test showed that accuracy for the first choices for emotions anger, fear, joy, and sadness is between 60 and 80%. For emotions surprise and disgust accuracy was lower then 60%. These results are comparable to subjective tests for other databases (Tickle, 2000). Only the first choice was considered in the analysis of Slovenian subjective evaluation. Majority of errors were made on short sentences and isolated words and few on long sentences and paragraph text.

7. Conclusion

The recorded emotional speech database represents a good basis for emotional speech analysis and is also usable for emotional speech synthesis. These databases present one of the largest databases of emotional speech

that are multilingual and based on common specifications. These databases could be also used for comparison of emotions between different languages.

Slovenian and Spanish subjective evaluation tests showed comparable results to other subjective tests of emotional speech databases (Tickle, 2000). Furthermore, the database is used to develop a multilingual emotion classifier and will be used for multilingual emotion modelling for speech synthesis.

The database is produced and owned by the different speech partners within the Interface project and has been sponsored by the European Commission in the scope of the Interface project IST-1999-10036.

8. References

Anscobe E. and Geach P. (1970). Descartes Philosophical Writing, *Nelson: The Open University*

Cornelius R. (1996). The Science of Emotion. *New Jersey: Prentice-Hall*..

Johnstone I. T. (1996). Emotional speech elicited using computer games. *Proceedings of the 4th International Conference on Spoken Language Processing*, 3, 1985-1988.

Kramer E., (1963). Judgment of Personal Characteristics and Emotions from Nonverbal Propertiesof Speech. *Psychological Bulletin,* 60.

Tickle A. (2000). English and Japanese Speakers' emotion vocalisation and recognition: A comparision highlighting vowel quality. *Proc. from ISCA Workshop on speech and emotion.*

AKG Acoustics inc., (1999), AKG 3000B Condenser microphone datasheet

Laryngograph Ltd., (1997), Portable Electro-laryngograph users manual