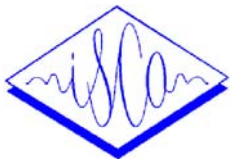

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

DESIGN, RECORDING AND VERIFICATION OF A DANISH EMOTIONAL SPEECH DATABASE

Inger S. Engberg, Anya V. Hansen, Ove Andersen and Paul Dalsgaard
Center for PersonKommunikation
Aalborg University, Frederik Bajers Vej 7 A2, 9220 Aalborg Øst, Denmark.
Tel. +45 96 35 86 78, FAX: +45 98 15 15 83, E-mail: ise@cpk.auc.dk

ABSTRACT

A database of recordings of **D**anish **E**motional **S**peech, DES, has been recorded and analysed. DES has been collected in order to evaluate how well the emotional state in emotional speech is identified by humans. The results sets a standard for identifying Danish emotional speech.

DES contains recordings from four actors, two of each gender. Actors were used for the recordings as they were believed to be able to realistically convey a number of emotions, namely: neutral, surprise, happiness, sadness and anger. The recordings from each actor consist of two isolated words, nine sentences and two passages. The complete database comprises approximately 30 minutes of speech.

A listening test with 20 listeners was conducted. The emotions were on the average identified correctly in 67,3% of the cases, with a [66,0 - 68,6] 95% confidence interval. An analysis reveals that most confusion occurred between surprise and happiness and between neutral and sadness.

1 INTRODUCTION

People with impaired speech have severe problems in their communication with other people, and to days existing speech synthesiser-based communication aids lack naturalness and flexibility in the representation of voices (female versus male) and are not able to convey an intended emotion [1].

One of the aims of the TIDE project **V**oices, **A**ttitudes and **E**motions in **S**peech **S**ynthesis, VAESS, [1] is to include a range of emotions in the synthesised speech. Before the synthesiser can be augmented with capabilities for conveying emotions, it is necessary to study these phenomena in daily life speech. This can e.g. be done by recording a database of speech displaying these phenomena, and submitting the recorded data to systematic analyses.

The design, recording and verification of such a database is presented in this paper. The results are discussed and compared with the results of a Swedish listening test described in [2].

2 DESIGN

The recordings of a database of emotional speech for Danish, DES, has been conducted under laboratory conditions, following a standard data recording procedure. This is necessary in order to systematically record the same utterance with different emotional contents. At the same time undistorted clean speech signals are recorded to serve as the basis for parameter analyses, which constitute the basis for the design of emotions for the formant based synthesiser, which is part of the communication aid used in the VAESS project.

2.1 Speakers

In some psychological experiments [3], real emotions have been elicited from subjects in a laboratory. This is, however, for ethical reasons not desirable. And as a consequence the emotional speech recorded with DES is spoken by a number of Danish actors from Aarhus Theatre [4].

It is shown in [5] that recordings with actors are good approximations to true emotional speech, justifying also their use with the present database. [5] compares recordings of a speaker reporting from a dramatic event with recordings of an actor simulating the reporter's emotional state during the event. Differences between the recordings were found, but in general the mode of speaking and the fundamental frequency range and variation were alike.

Four actors familiar with radio theatre, were employed for the recording of DES, see Table I for gender and age of the actors.

2.2 Prompting Text and Emotions

To avoid that a listeners decision about the emotional contents of a sentence is coloured by its semantically meaning, it has been attempted to construct semantically neutral sentences.

In order to analyse the emotional content in both very short and longer utterances, it was decided to record two short single words (yes and no), nine sentences (four of which were questions) and two passages of fluent speech. The prompting text can be

Initials	Gender	Age
DHC	Female	34
KLA	Female	52
JZB	Male	38
HO	Male	52

Table I. *Gender and age for the four actors used in the recording of DES.*

found in the documentation of DES [6]. Each utterance was spoken with each of the five emotions in mind: neutral, surprise, happiness, sadness and anger.

3 RECORDING

DES was recorded in an acoustically damped sound studio at Aarhus Theatre [6]. The studio was "floating" on rubber absorbers as is the operator room. Between the studio and the operator room an angled window with three layers of glass (of different thickness) was placed so that the actors and the operators had visual contact all the time. The operators could get in contact with the actor via an intercommunication system. In addition, the operators were continuously listening to the actors via the recording chain, and could if necessary, interrupt the actor. Speech from each of the actors were recorded in separate sessions. This prevented the actors from influencing each others speaking style.

The session was started by giving the actors a thorough description of the experiment, and they were asked for questions regarding pronunciation and emotions. The actors were placed at a table with their arms on the table in order to keep a constant distance to the microphone during the whole recording. The actors were urged to ask, at any time, for a break. Prior to the recordings, the actors had been given a very brief introduction to the recordings and the prompting text. The actors were asked to use their own every day way of expressing emotional states, and not the exaggerated emotional expression known from stage acting. All utterances were spoken with one emotion at a time.

For the recordings a high quality microphone was used. In a mixer a 5 Hz high pass filter cut of the lowest frequencies before the whole session was recorded on a DAT tape at 48 kHz sampling frequency. The DAT tape recorder was started when the gain was set and was only stopped when the actor had a break or left the studio. In this way not only the required recordings was recorded but also the communication between the operators and the actor, and mistakes by the subjects.

4. VERIFICATION

Following the recordings, a listening test was performed to test whether normal listeners could identify the type of emotion with which the utterances had been recorded. 20 subjects (10 of each gender and mainly staff at CPK) were used. They had an average age of 38 years, ranging from 18 to 59 years, see Table II.

Age	Male	Female
Under 20	PED	IHH
21 - 30	SVA, PLE	AVH, HC, JFV
31 - 40	OA, JPM, PE	BLR, VJP
41 - 50	HEB, POR, SPJ	JT
51 - 60	HE	ILW, GE, JEB

Table II. *Ages and gender of the 20 listeners.*

Four listening tests, one for each of the actors, were prepared on a DAT-tape. Each listening test presented 13 (2+9+2) utterances spoken with the five different emotions. All the listeners went through the four tests. Accumulated the listening test lasted from one hour to nearly four hours (the longer sessions were spread out over two days), but most listeners took about two hours incl. breaks.

No training session was offered to the listening subjects before the test and they were not given any feedback during the session. The listeners were asked to judge the emotional contents of the utterance by a forced choice. The listeners were allowed to hear the utterance several times before deciding on the emotional category. They were, however, neither allowed to go back to compare with earlier utterances nor to change an earlier choice. The listening test did not always start with the same actor, but the succession of the actors was the same. The listeners were free to ask for breaks when needed, but a break was made after each actor.

After each of the four listening test, the subjects were asked to state whether they found the task of identifying emotions *very easy*, *easy*, *neither easy nor difficult*, *difficult* or *very difficult*. They were further asked to describe the factors if any that made them choose the different emotions. Finally they were asked for further comments they might have to the listening test of this particular speaker.

5. RESULTS

The results of the verification session are given in Table III. In each row the % are calculated on the basic of 1040 utterances of that emotion. Total shows how often an emotion was selected by the listeners, and is calculated on the basic of all 5200 utterances. The emotions were correctly identified on the average in 67,3% of the cases, with a [66,0 - 68,6] a 95% confidence interval, ranging from one listener who had an average of 55,4% to another who had 80,4%. It was a characteristic for the tests, that surprise often was confused with happiness as well as neutral and sadness, but to a lesser degree.

From the scores in Table III it can be seen that, the emotion "sad" is chosen by 24,3% which contributes to a high identification rate on 85,2%. The emotion "angry" was only chosen in 17% of the presentations,

Listeners ⇒ Actors ⇓		RESPONSE in %				
		Neu.	Sur.	Hap.	Sad.	Ang.
STIMULUS	Neu.	**60,8** 57,8-63,7	**2,6** 1,8-3,8	**0,1** 0,0-0,6	**31,7** 29,0-34,6	**4,8** 03,7-06,3
	Sur.	**10,0** 8,3-12,0	**59,1** 56,1-62,1	**28,7** 26,0-31,5	**1,0** 0,5-1,8	**1,3** 0,7-2,1
	Hap.	**8,3** 6,8-10,1	**29,8** 27,1-32,7	**56,4** 53,4-59,4	**1,7** 1,1-2,7	**3,8** 2,8-5,1
	Sad.	**12,6** 10,7-14,8	**1,8** 1,2-2,8	**0,1** 0,02-0,6	**85,2** 82,9-87,2	**0,3** 0,1-0,9
	Ang.	**10,2** 8,5-12,2	**8,5** 6,9-10,3	**4,5** 3,4-6,0	**1,7** 1,1-2,7	**75,1** 72,4-77,6
Total		**20,4** 19,3-21,5	**20,4** 19,3-21,5	**18,0** 16,9-19,0	**24,3** 23,1-25,5	**17,0** 16,0-18,1

Table III. *Confusions between the emotions for all speakers and listeners with coherent 95% confidence interval. Neu. is short for neutral, Sur. for surprise, Hap. for happy, Sad. for sadness and Ang. for angry. Total shows how often the different emotions were chosen by the listeners.*

but when presented to the listeners it was identified correctly in 75,1% of the tests.

Since the listeners were not offered a training session before the listening test, it was tested whether the listeners scored higher on the 20 last presentations than on the first 20. A detailed analysis shows that 62,6% of the first 20 presentations were perceived correctly with a [60,2 - 64,9] 95% confidence interval, whereas the score for the last 20 utterances is 73,2% with a [71,0 - 75,3] 95% confidence interval. However not necessarily all emotions were represented in the first/last 20 presentations.

In identifying emotional categories it was further tested whether females were better than males. The results show that the females perceived the correct emotions in 68,8% of the presentations correctly with a [67,0 - 70,6] 95% confidence interval whereas the males perceived the correct emotions in 65,9% of cases correctly with a [64,1 - 67,7] 95% confidence interval.

Utterance	Correct	95% Confidence Interval
Single Words	67,5%	64,2 - 70,7 %
Sentences	65,3%	62,3 - 66,9 %
Passages	76,3%	73,2 - 79,1 %

Table IV. *The scores for the different utterances with coherent 95% confidence interval.*

DES contains both single words, sentences and passages. It was tested whether their was a difference in their sores. In Table IV it can be seen that passages was easiest to identify emotions from.

The experiments showed that it is very difficult to distinguish between questions and surprised sentences. Four of the nine sentences in DES were questions. The sentences spoken with the surprised emotion were parted in Q and N:

- Q: containing the 1600 inquiring sentences, of which 320 were surprised.
- N: containing the 2000 non inquiring sentences, of which 400 were surprised.

The percentage of sentences interpreted as surprised relative to the number of surprised sentences was then calculated for each group. In group Q 168,4% was interpreted as surprised contrasting to 67,0% in group N.

The listeners impression of the difficulty of the test together with the actual score for the different actors can be seen in Table V. Not all listeners answered this question for each actor, hence total in Table V is not always 20. 75% of the listeners found it *difficult* or *neither easy nor difficult* to identify the emotions. The actors DHC and JZB were in general found from *difficult* to *easy*, where as HO and KLA were characterised as *very difficult* to *neither easy nor difficult*, except for one listener.

Initials	HO	DHC	JZB	KLA	Total
Score	62,4%	68,3%	72,1%	66,5%	67,3%
very difficult	3	1	-	3	7
difficult	11	4	4	10	29
neither / nor	5	9	8	6	28
easy	-	5	5	1	11
very easy	-	-	1	-	1
Total	19	19	18	20	76

Table V. *The judged difficulty together with the actual score for the different actors. "Neither / nor" stands for "neither easy nor difficult", and a "-" means that no one chose this option.*

In the last remarks, that the listeners were asked to give about each of the listening tests, it can be seen that some listeners found HO's emotional state difficult to hear due to his deep voice Also the fact that DHC was younger than KLA was stated by a listener. Many listeners stated, that they had chosen the neutral emotion, when no other emotion seemed to be present.

6. DISCUSSION

In a Swedish listening test described in [2], six emotional categories were correctly identified in 81% of the cases. However the Swedish listening test consisted of only 72 sentences, selected from six sentences spoken with six different emotions more than once by two actors

(one of each gender). In the Swedish listening test set all emotions were presented 12 times, but the same sentence spoken by the same actor with the same emotions could be present more than once.

The Danish listening test set included only five emotional categories which on the average were correctly identified in 67% of the cases. In the Danish listening test all utterances spoken with all emotions by all actors were tested. An utterance spoken with one emotion by one actor is not present more than once.

The results of the Danish listening test are not as good as the Swedish results. In the Swedish listening test no neutral emotion was present as it was in the Danish. In the remarks made by the listeners it was often stated that the neutral emotion was chosen when no other emotion seemed to be present - as a sort of garbage can. In Table III it can be seen that neutral has scored 10% when all other emotions were presented. Indicating that the neutral emotion was used as a garbage can, and thereby negatively influenced the score of the Danish listening test.

7. CONCLUSION

The emotional categories were correctly identified on the average in 67,3% of the cases, ranging from 55,4% correctly to another listener identifying 80,4% correctly.

The evaluation of DES showed that neutral and sadness were confused. The neutral emotion was chosen when no other emotion was clearly present. The sad emotion, on the other hand, shows a strong confusion with neutral. In fact 31,7% of the actors neutral speech were identified as being sad by the listeners. From the listening test it is seen that neutral Danish speech often is perceived as sad.

A training session was not given prior to the listening test. There was a difference of 10% between the score of the first 20 utterances and the last 20, showing that the listeners had adopted to the voice in question.

It was tested whether female and male listeners performed equally well when identifying emotions, and with a 95% significance interval, there can be found no difference in their performance in this test [7].

The type of utterance was found important. The passages were easiest to identify emotions from. This could be because there is most data in the passages, as they are the longest. Between the single words and the sentences there were little difference.

Four of the nine sentences were questions. From the test it can be seen that the inquiring sentences often were interpreted as surprised even though another emotion was in question. In future, questions should not be used as semantic neutral sentences.

The listeners found it easiest to judge the emotional speech from JZB, who also got the highest score. All in all the impression of the difficulties corresponded well with the actual score. In general the emotions of the two younger actors were judged to be easier to identify than the emotions of the older actors. This indicates that emotions expressed by younger people perhaps are easier to identify.

The results presented in this paper show that in general it is difficult to identify emotions from real emotional Danish speech. However the “sad” and the “angry” emotion were identified in 85 and 75% of the cases, and work on including these two emotions into Danish synthesised speech has been started.

7.1 Availability of DES

DES is collected only for scientific use, and is available for such on a CD-ROM. DES has been phonotypical transcribed using SAM-PA [8], which is included on the CD-ROM together with documentation of DES. The CD-ROM can be acquired, for a minimal handling and postage fee, by contacting CPK.

8. ACKNOWLEDGEMENTS

The VAESS Tide Project was funded by the EU and was a collaboration between KTH in Sweden, GTH in Spain, Sheffield University and BiDesign in UK and CPK in Denmark. The assistance and support of the co-operating partners is gratefully acknowledged. The Danish Technical Research Council has supported the research reported in this paper by its financial support to CPK.

9. REFERENCES

[1] Tide Project : TP1174 - VAESS, Technical Annex, 26-6-1995.

[2] Öster, Anne-Marie & Risberg, Arne (1986) The Identification of the Mood of a Speaker by Hearing Impaired Listeners, Speech Transmission Lab. - Quarterly Progress and Status Report 4 1986.

[3] Murray & Arnott (1992) Towards the Simulation of Emotion in Synthetic Speech: A review of the literature on human vocal emotion, Journal of Acoustic Soc. Am. Vol. 93, No. 2, Feb. 1993, pp1097-1108.

[4] Aarhus Teaters Lydstudie, Skolegade 9, 3. sal, 8000 Århus C, Denmark.

[5] Williams, C.E & Stevens, K.L (1972) Emotions and speech: Some acoustical correlates, Journal of Acoustic Soc. Am. Vol. 52, No.4, part 2, pp. 1238-1250.

[6] Documentation of the Danish Emotional Speech Database DES, Inger S Engberg & Anya V. Hansen, 1997, Center for PersonKommunikation, Fredrik Bajers Vej 7A-2, Aalborg University, DK-9220 Aalborg.

[7] Ross, Sheldon M (1987), Introduction to Probability and Statistics for Engineers and Scientists, John Wieley & Sons, Inc.

[8] J. C. Wells, “Computer-coded Phonemic Notation of Individual Languages of the European Community” Journal of the International Phonetic Association (1989) 19(**1**), 31-54.