Designing and Recording Emotional Speech Databases

Swati D. Bhutekar
S.R.K.N.E.C
Nagpur, India

M. B. Chandak
S.R.K.N.E.C
Nagpur, India

ABSTRACT
This paper describes the factors used in designing and recording large speech databases for applications requiring speech synthesis. Given the growing demand for customized and domain specific voices for use in corpus based synthesis systems, good practices should be established for the creation of these databases which are a key factor in the quality of the resulting speech synthesizer. This paper focuses on the factors affecting to the designing of the recording prompts, on the speaker selection procedure, on the recording setup and on the quality control of the resulting database. One way to find the emotions in the speech is , Once the speech has been recorded from the user it is converted into text, at the same time the stressed word from the speech is recorded & then the frequency for that word is find out for recording the corresponding emotion.

General Terms
Natural Language Processing

Keywords
Extraction of Emotion in Speech, Database, Recording.

1. INTRODUCTION
There are currently several speech synthesis systems with enough naturalness to be used in applications for a wide range of domains. The use of these systems has been restricted by the number of available voices and by the occurrence of artifacts when synthesizing less common words. Companies do not like to have an interactive voice response system (IVR) with the acme voice as theirs competitors and most applications of speech synthesis require the use of domain specific words, like brand names or technical terms with unusual phonetic sequences. These restrictions are a consequence of the technology used in most speech synthesizers that are based on the concatenation of variable length speech units taken from an inventory of recordings of a single speaker. To assure the coverage of the most common sequences in a given language, the inventory must contain a considerable amount of speech recordings (from 3 to 10 hours or more) with carefully selected contents. These contents are designed to provide a good coverage of the phonetics and intonation of the selected language using analysis performed on available text corpora, mainly newspaper texts and books that do not always cover the specific requirements of certain applications such as speech-to-speech translation, medical systems, customer support, etc. Also, the recording of the inventory requires a large number of recording sessions and a strict recording procedure to assure the uniformity of the database. The high cost of the recording process limits the ability of the technology providers to produce more than a few voices for each language. A solution to this problem has been to separate the speech synthesizer engine from the inventory that defines the synthesizer's voice. Several of the public available systems allow the integration of new voices, like Festival and MBROLA, and some companies are willing to outsource the recording procedure in exchange for wider range of customers. Also, a properly recorded voice can be used in several systems using different technologies and have a lifespan longer than the synthesizer engines.

This paper describes the various factors for designing and recording large speech databases for applications requiring speech synthesis.

2. FACTORS NEED TO BE CONSIDERED, REQUIREMENTS FOR EMOTION MODELLING
The study of the various models of the emotions requires recording samples of all the basic emotions. This paper considers the set of emotions known as "the Big Six" (Cowie & Cornelius, 2003): sadness, happiness, anger, fear, surprise and disgust. Additionally, neutral style has also been considered. Different types of corpora have been used for the study of emotions in speech. Some groups have employed spontaneous emotional speech, trying to get the greatest authenticity in the emotions. Others have worked with elicited emotions, putting the speaker into situations to rouse a specific emotion. A third option has been to use a speaker with acting skills to simulate emotions. Though this latter technique can exaggerate the emotions, the fact is that they are recognized, so that practical modelling can be derived from them.

Speakers' variability

Previous works have also taught us that it was impossible for the speakers to keep a constant reference level for their rhythm, tone, volume, etc. through a long lasting recording session. The expected recording time for the database spread through several sessions, so the effects of these variations were supposed to be even more important. In order to quantify these deviations and keep on being able of comparing prosodic parameters among emotions, a control text (**short continuous text**) can also designed which can be read with neutral style at the beginning, mid-session and end of every session. In this way, the reference levels in the prosodic parameters for each session will be extracted from this control text, and the data of every emotion will be normalized against these reference levels.

2.1 Speaker Selection
Usually a speaker for a unit selection voice has to be able to speak in the same tone of voice constantly for long periods of time. In the case of building an emotional unit selection voice, the speaker should be able to portray an emotion for long periods of time. Both attributes are important for recording a

emotional database. To determine the quality of a speaker, auditions were held. The following paragraphs describe the aspects that were looked at in determining which speaker to chose. During the audition, the speakers had to read 10 sentences of about equal length that were printed on one single sided page. They were asked to read the sentences three times; once in a neutral voice, once in an angry voice, and once in a happy voice. They always read the neutral part first, then they could decide if they wanted to read either the angry part or the happy part next.

Recordings were made of the auditions in the studio where the actual database collection would take place. This was to see if the speaker felt comfortable in a studio setting. The most important factor in determining a speaker's quality was his or her reading abilities. A speaker would need to read for a few hours, preferably without mistakes. It is very tiring on the speaker if he had to repeat sentences because he made a mistake.It was also important that the researcher liked the voice. He would have to listen to that voice for extensive periods of time during the voice building process. Another important factor was the quality of the voice. Since many aspects of the voice building process were automated, a "cleaner voice" would get better results. A breathy voice would not work as well with the tools provided. The previous attributes of a speaker apply to all unit selection voices, but in the case of emotional speech synthesis, it was also important that the speaker was able to portray emotions convince-ably. It is not clear if one should use actors to portray emotions or if amateurs are convincing enough. Both approaches were tried in this MSc thesis. Ideally a number of people would have rated the recordings of the auditions according to the emotions portrayed to see how good the speakers were. But in this case the recordings were only listened to by the researchers to decide which speaker was the best.

The decision criteria for speakers :
• the reading naturalness;
• the duration of the recording session (number of repetitions);
• the capability of maintaining the voice quality during the recording session;
• the pleasantness of the synthesized voice;
• the voice ability to mask concatenations errors.

2.2 Professional Actor vs. Regular Speaker

During the voice building process, it became very clear that trained speakers are more suitable for collecting an emotional speech database. The actress was more comfortable in the studio setting and therefore made less reading mistakes. This is reflected in the time difference for the recording sessions. The non-professional speaker needed about twice as much time for 400 sentences than the female speaker due to reading errors.Conveying the intended emotion to the non-professional speaker was very difficult. The approach taken was to play him emotional speech to let him imitate the voice. Some sessions sounded more convincing than others. In general he was very good in portraying the angry voice and not as good at portraying the happy voice. The speaking rate of the angry voice was very fast which made it harder to build the voice. The speaker told me that at points, he really felt a certain emotion which was reflected in the quality of the performance. Also, it was very hard for him to read sentences in a certain emotion if the connotation of the sentence was suggesting a different emotion. Since the script was made out of newspaper sentences, there were a few sentences about war. The speaker made noticeably more mistakes when he had to read these sentences in a happy voice.

The actress was able to speak with the same speaking rate for extended periods of time while making about 1 or 2 mistakes every 50 sentences. It was also straightforward to convey the intended emotion to her. She was told to read a session in a certain voice at a certain speaking rate, which was enough information for her to be able to convey the emotion convince-ably. The actress did not have any problems with the content of the sentences. In general, her angry voice was better than her happy voice. There are several advantages in using an actor to portray emotions. The recording sessions with the actress were a lot quicker and her session waveforms sounded a lot more "professional". But it is unclear if her portrayal of emotions was natural. The recordings were clearly understandable as being in a certain emotion but they were not necessarily natural. Most of the recordings of the regular speaker did not sound very emotional, but when they did, it sounded natural. It seemed he really felt that emotion where as the actress' portrayal did not convey the same feeling.

2.3 Phonetic balance

Besides assuring that large units will be found in the database, it is also necessary to assure that all the possible phonemes and certain phoneme combinations of a language are included. If we want to assure that the unit selection synthesis will produce at least the quality of other concatenative methods, we will have to design the database to guarantee that there are at least all the smallest units used by these other methods. A reasonable minimum size for these units is diphoneme. Once the minimum unit is selected, the purpose of the phonetic balance is to keep the appearance rate of these units in the database corpus, as close as possible to their appearance in the actual language. In this way, usual diphonemes will appear lot of times in the recorded database, in multiple contexts, and rare ones will appear perhaps only once, or even they will have to be explicitly added. In addition, there are some problematic combinations of three or four phonemes, so some units longer than diphonemes, called "poliphonemes" have to be considered. 406 of these poliphonemes have been defined for Basque.

2.4 Evaluation of the natural voice

The assessment of the natural voice is aimed at judging the appropriateness of the recordings as a model for readily recognisable emotional synthesised speech.

The voice quality was tested by non-handicapped listeners, since their perception would not be any different . fifteen normal listeners, both men& women of different ages were selected from several social environments; none of them was used to synthetic speech.

The stimuli contained 5 emotionally neutral sentences. 3 sentences came from the short sentences set & the other 2 were part of the passages. As 3 emotions & a neutral voice had to be evaluated, 20 different recordings per listener & session were used.

In each session the audio recordings of the stimuli were presented to the listener in a random way. Each piece of text was played up to 3 times.

2.5 Customizable Parameters of emotional synthesis

Speaking rate: ranging from 150 for a sad voice to 179 for an angry one; the default value for a neutral voice is 160.

F0 Range; sad voice had the smallest range & angry voice had the highest one (hot anger).

Mean Pitch level: a happy voice had the highest mean F0; a sad voice had the lowest one.

F0 slope: angry & happy voices shared a high descendent slope; sad voice was rather flat.

Spectral Tilt**:** a lower tilt value increases the high frequency contents of the voice source, producing clearer voices; it is a specially useful parameter for happy voices.

Additive Noise**:** pitch-synchronous noise added to the voice source; used for sadness & anger.

Type of emotion: special rules are applied for happy & neutral voice, changing the default intonation contour characteristics.

Table 1 : The frequencies of phonemes in the corpuses for each language.

English	
Sample	Freq. (%)
A	2,11
E	3,30
{	11,18
OI	1,41
@U	0,79
I@	1,84
b	1,93
C	0,49
d	4,12
D	2,75
E	2,14
F	1,69
G	1,03
h	1,47
i	4,69
I	3,39
K	3,57
l	3,84
m	2,92
n	7,24
N	1,10
O	1,35
o	0,08
p	2,22
r	4,54
s	5,30
S	0,69
T	0,32
t	7,05
u	0,38
v	2,00
w	2,13
z	3,29
Z	0,04

2.6 Database Selection

2.1.1 Database size

The size of the database has to be carefully fixed in order to assure that, at synthesis time, units as large as possible are found. Appropriate size starts from 1 hour of recordings (Febrer, 2001). This means approximately 40,000 diphonemes, which translated into Basque words (with an average of 6.3 diphonemes per word) yields to some 6,400 words, or 500 phrases.These figures establish the bottom limit in the database size. The final size will be influenced by the other requirements. Obviously, the bigger the database is, the better the synthesis results could be, but there are performance, resources consumption and even speaker availability constraints that set the upper limit not very far away from the minimum one.

2.1.2 English speech emotion databases

Database 1. The database was recorded at the Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia . It contains emotional speech in six emotion categories, such as disgust, surprise, joy, fear, anger and sadness. Two neutral emotions were also included: fast loud and low soft. It is seen that the emotion categories are compliant with MPEG-4 . Four languages (i.e. English, Slovenian, French and Spanish) were used in all speech recordings. The database contains 186 utterances per emotion category. These utterances are divided in isolated words, sentences both affirmative and interrogative, and a passage.

Database 2**.** R. Cowie and E. Cowie constructed this database at the Queen's University of Belfast. It contains emotional speech in 5 emotional states: anger, sadness, happiness, fear and neutral. The readers are 40 volunteers (20 female, 20 male) aged between 18 to 69 years. The subjects read 5 passages of 7-8 sentences written in an appropriate emotional tone and content for each emotional state. Each passage has strong relationship with the corresponded emotional state.

Database 3**.** Belfast Natural Database. R. Cowie and M. Schroder constructed this database at Queen's University. The database is designed to sample genuine emotional states and to allow exploration of the emotions through time. Two kinds of recordings took place. One was recorded in studio and the other direct from TV programs. A total of 239 clips (10-60 sec) is included in the database. The clip length is taken to be quite long in order to reveal the development of emotion through time. The studio recordings consist of two parts. The first part contains conversations between students on topics, which provoke strong feelings. The second part contains audio-visual recordings of interviews (one-to-one) involving a researcher with fieldwork experience and a series of friends.

Database 4**,** Kids' Audio Speech Corpus NSF/ITR Reading Project. R. Cole and his assistants at the University of Colorado recorded database 4. The aim of the project was to collect sufficient audio and video data from kids in order to enable the development of uditory and visual recognition systems, which enable face-to-face conversational interaction with electronic teachers. The Kids' Audio speech Corpus is not clearly oriented to elicit emotions. Only 1000 out of 45000 utterances are emotion oriented.

Database 5**.** Emotional Prosody Speech and Transcripts. M. Liberman, Kelly Davis, and Murray Grossman at the University of Pennsylvania constructed database 5 . The

database consists of 9 hours of speech data. It contains speech in 15 emotional categories, such as hot anger, cold anger, panic-anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt.(there are six more)

3. PERSPECTIVES ON EMOTION

There are four basic traditions in emotion research in Psychology. Each theory focusing on different components and making different assumptions on what is important for describing an emotion.

3.1 The Darwinian perspective

Charles Darwin in his book ***"The Expression of Emotion in Man and Animals"*** laid the groundwork for much of modern psychology and also for emotion research. He describes emotions as reaction patterns that were shaped by evolution. This implies that emotions are common in all human beings and also that some emotions might be shared with other animals. The concept of basic emotions was also developed by Darwin. The function of the emotions may be a biological activation to make an animal more responsive to certain situations, including the tendency to perform certain actions. Another function might be a signal to an external observer, such as threat and therefore influencing the observers behaviour. The universality of facial expression as found by Ekman was an important finding in support of the Darwinian view. It makes clear that emotions have a biological basis and are therefore evolutionarily shaped. He demonstrated at least six emotions (happiness, sadness, anger, fear, surprise, and disgust) that were expressed in the face and recognised in the same way in many cultures.

3.2. The Jamesian perspective

For William James the body is essential for an emotion. Bodily changes follow some stimulus automatically and the emotion arises through the perception of these changes. Therefore without the perception of the body there are no emotions. The facial feedback hypothesis follows the Jamesian perspective. It states that the facial expression of a person has an effect on the subjective emotional experience. For example if a person has a facial muscle configuration corresponding to a happy face the person reports feeling happier (e.g. smiling makes one happy).

3.3. The cognitive perspective

Cognitive emotion theories relate emotions to appraisal, which is the automatic evaluation of stimuli by low level cognitive processes. It determines how important a given stimulus is for the individual and increases the chances of an appropriate response. Scherer's component process model makes physiological predictions relevant to speech from such appraisal processes. The model details the appraisal process as a series of stimulus evaluation checks (SEC) in a certain temporal order: novelty check, intrinsic pleasantness check, goal/need significance check, coping potential check, and norm/self compatibility check. Each SEC is associated with an appropriate response in the various components of emotions (see Componential View). An emotion is denoted in the component process model as a configuration of SEC outcomes.

3.4. The social constructivist perspective

In the social constructivist perspective, emotions are seen as socially constructed patterns that are learned and culturally shared . Emotions have a social purpose that regulates the interaction between people. The expressions of emotions and the emotions themselves are described as culturally constructed. Although the biological basis of emotions is recognised, the socially constructed mechanisms are given more weight.

4. THE APPROACHES OF ESTIMATION OF EMOTION IN SPEECH

Following are the three methods for estimation of emotions in speech:

Using speech synthesize, Adding SD, skewness and kurtosis to the statistical value of features and Using the classifier for each emotion. These methods are explained in more detail in this section.

4.1 Using Speech Synthesize

We need human speech data in a conventional method. It is a hard work for researchers. For example, human speech includes noise. Researchers need a lot of time to collect speech. And, it is difficult for most examinee to express certain emotion like as actors or actress. In order to reduce the load of this work, we use speech synthesize. We can use some speech synthesize software like Microsoft Speech SDK and Festival . So it is not difficult to synthesize various speeches. Most software, however, do not have an emotion expression facility. But, in order to make a classifier, we need to know emotion expressed in each speech. So, we put emotion labels on synthetic speeches based on evaluation by people using the following method. We synthesize some speeches from same phrase. But, parameters like pitch, speed and volume are different each other. People estimate emotion in speech and answer it. If an emotion is estimated by more than half of judges, we put the emotion label on the speech. There are some strong points in this method. One is that data does not include noise. The other is that evaluating synthetic speech is easier than expressing the certain emotion in speech. In addition, it is not difficult for researchers to collect many data.

4.2. Using the classifier for each emotion

We use only one classifier for every emotion in conventional method. This classifier tries to estimate one of emotions. We think that people express some emotions in speech at once. speech features mix the feature based on each emotion. The relation between features and emotion is very complex. It is difficult to estimate emotion using one classifier. So we makeThe classifier for each emotion. Each classifier is specialized one emotion and shows whether there is the emotion in human speech. We regard the set of prediction values gotten by classifiers as emotion in speech. When we want to estimate only one emotion in speech, we estimate emotion which has the maximum value of the prediction value. If we ca not select one emotion based on the prediction value, we do not estimate emotion in speech.

5. CIRCUMPLEX MODEL OF AFFECT

Instead of independent emotion categories, several researchers have described an effective space . Russell has developed a circular ordering of emotion categories that makes it straightforward to classify an emotion as close or distant from another one. By having subjects rate similarity of different emotion words and converting the ratings into angles, a circular ordering emerged.

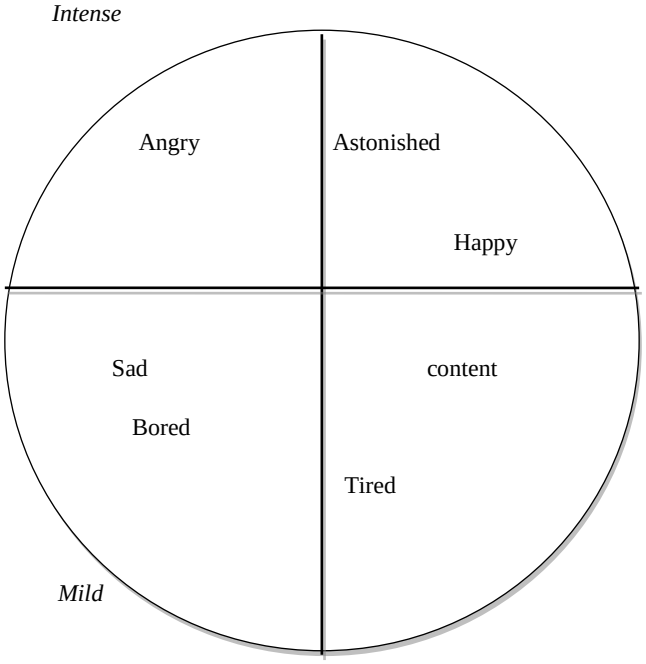

Figure 1: Circumplex Model of Affect as described by Russell (1980)

In addition to the circular ordering, Russell found evidence of two dimensions d describing effect . He called the two dimensions ”valence” and ”arousal”. These terms correspond to a positive/negative dimension and an activity dimension respectively. Scherer found further evidence supporting two dimensions but proposed a different interpretation relevant to the component process model .Some researchers suggested that using emotion dimensions gives an impoverished description of emotions . Depending on the application, the use of emotion dimensions is often sufficient, sometimes even necessary. Especially the ability to measure sssimilarity becomes very useful in computing applications where distance measures are used. The dictionary of affect that is used in the practical part of this MSc dissertation makes use of emotion dimensions.

6. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the emotional speech synthesis.

.

7. REFERENCES

[1] Gregor O. Hofer **,** “Emotional Speech Synthesis”, Master of Science School of Informatics University of Edinburgh 2004

[2] Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Iker Luengo, Aholab - “Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque”, Dept. of Electronics and Telecommunications. Faculty of Engineering. University of the Basque Country.

[3] Inger S. Engberg, Anya V. Hansen, Ove Andersen and Paul Dalsgaard**,** “Design, Recording and verification of a Danish Emotional speech Database”

[4] Lu´ıs C. Oliveira, S´ergio Paulo, Lu´ıs Figueira, Carlos Mendes, Ana Nunes‡, Joaquim Godinho‡ ,“Methodologies for Designing and Recording Speech Databases for Corpus Based Synthesis”

[5] Masaki Kurematsu, Jun Hakura and Hamido Fujita, “An Extraction of Emotion in Human Speech Using Speech Synthesize and Classifiers for Each Emotion”, in International Journal of Circuits, Systems and Signal Processing

[6] Dimitrios Ververidis and Constantine Kotropoulos, “A Review of Emotional Speech Databases”, in Proc. 9th Panhellinic Conference on Informatics (PCI) , pp-560-574,Thessaloniki, Greece, November 2003

[7] Voice", Irvine, **“**Models of Speech Synthesis**”,** draft version of a paper presented at the "Colloquium on Human-Machine Communication California, February 8-9, 1993, organized by the National Academy of Sciences, USA.

[8] “Features and Algorithms for the Recognition of Emotions in Speech”, in Proceedings of the 1st International Conference on Speech Prosody (2002)

[9] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE transaction on speech and audio processing, vol.13, 2005**.**

[10] B. Kort, R. Reilly, and R. W. Picard, "An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion.," presented at In Proceedings of International Conference on Advanced Learning Technologies (ICALT 2001), Madison, Wisconsin, August 2001.

[11] Slobodan T. Jovičić, Zorka Kašić, Miodrag Đorđević, Mirjana Rajković, “ Serbian emotional speech database: design, processing and evaluation”, presented at SPECOM’2004: 9th Conference Speech and Computer St.Petersburg, Russia September20-22,2004