EMOTIONAL SPEECH SYNTHESIS: APPLICATIONS, HISTORY AND POSSIBLE FUTURE

Felix Burkhardt, Joachim Stegmann

Deutsche Telekom Laboratories
Berlin, Germany

Abstract: Emotional speech synthesis is an important part of the puzzle on the long way to human-like artificial human-machine interaction. During the way, lots of stations like emotional audio messages or believable characters in gaming will be reached. This paper discusses technical aspects of emotional speech synthesis, shows practical applications based on a higher level framework and highlights new developments concerning the realization of affective speech with non-uniform unit selection based synthesis and voice transformation techniques.

1 Introduction

No one ever speaks without emotion. Despite this fact, emotional simulation is not yet a self evident feature in current speech synthesizers. One reason for this lies certainly in the complexity of human vocal expression: current state-of-the-art synthesizers still struggle with the challenge to generate understandable and natural sounding speech, although the latter demand already indicates the importance of affective expression.

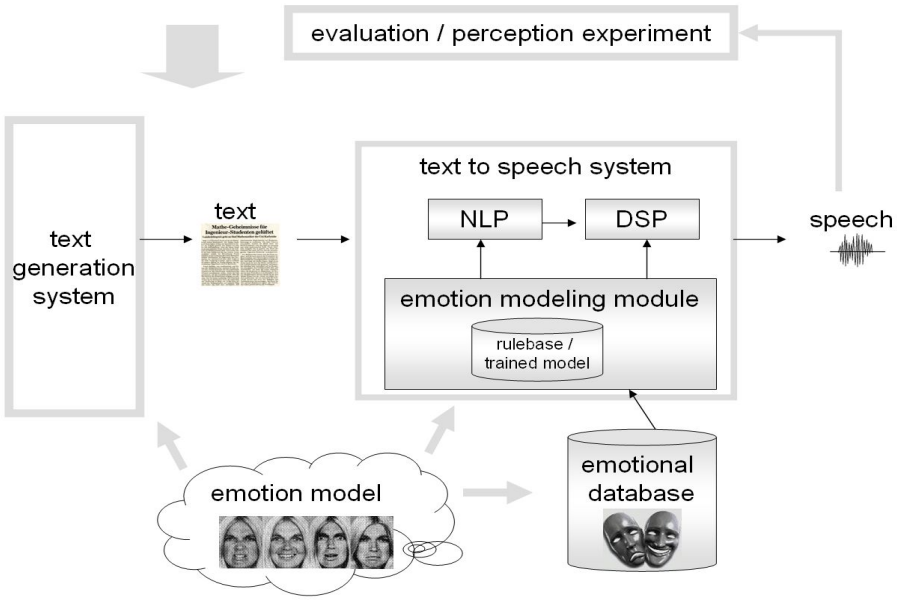

Figure 1 - General architecture of an emotional Text to Speech system.

This article is meant to give an overview on the state of art with respect to different aspects of emotional speech synthesis, ranging from use cases and emotion models to technical approaches.

The topic of emotional behavior is a most imprecise and difficult issue to describe. As a pair of psychologists famously remarked: “everyone except a psychologist knows what an emotion is”.

Fortunately, this definition problem concerns emotion recognition much stronger than emotion synthesis, because in the first case one needs an accurate model of "the real world" in order to capture the multitude of emotonal expression, while in the latter, a simple model based on a few basic emotions might be sufficient. Simulated emotional expression gets very well recognized when reduced to the display of some exaggerated prototypical emotions ([1, 2, 3]).

Generally spoken, speech synthesis can be divided between the following three kinds (although hybrid forms are possible).

- Voice response systems like used in an underground announcement system,
- Re- or copy-synthesis is used to alter a speech signal in its voice-related features. A special case would be voice transformation, for example changing in the case of voice conversion, a speech signal from a source to a target speaker. Voice transformation techniques can be used to generate an emotional expression.
- Arbitrary speech synthesizers in contrast can process any kind of input, given the limits of a target language. It must be noted though that all synthesizers are somewhat domain restricted. They might be divided into text-to-speech vs. concept-to-speech systems, depending on the information given with the text. With respect to the topic at hand, concept-to-speech systems might be able to label text automatically with target emotions.

An overall architecture of an emotional speech synthesis system is shown in Figure 1. The text to be synthesized, is either given as input or generated by a text generation system. Although text generation systems are out of scope of this article, formalisms to annotate text emotionally get discussed in section 4.

A text-to-speech synthesizer converts text into speech by first analyzing the text by natural language processing (NLP) and conversion in a phonemic representation aligned with a prosodic structure, which is passed to a digital speech processing (DSP) component in order to generate a speech signal. Both of these sub modules might be influenced by the emotion modeling component. Approaches to generate emotional speech will be discussed further in section 6.

Features of the speech signal are spectral (the "sound" of the voice), prosodic (the "melody" of the speech), phonetic (kind of spoken phones, reductions and elaborations), ideolectal (choice of words) and semantic features. All of these get influenced by emotional expression, although, for practical reasons, speech synthesis systems must refrain to take only a subset of this features into account. Aspects of emotional expression affecting speech features are discussed in section 5.

The component responsible to generate the emotional expression, needs to be trained on a database of emotional examples, irrespective of whether rules are derived from the data or statistical algorithms get trained. Databases are often recorded by actors [4], taken from real life data [5], or TV-data is used [6].

From these, a rule base or model can be generated and used to control the synthesizer. All the components of the emotion processing system need to have the same emotion model as a basis. Different approaches to designate and describe emotions are discussed further in section 3.

If a system is up and running, its success can only be judged by human listeners. Performance results and listening test designs will be discussed in section 7.

The quality of an emotional synthesizer of course depends primarily on its application: a synthesizer giving cartoon figures a voice meets different demands than a system to make the voice of a speech disabled person more natural. Applications of emotional speech synthesis will be discussed in the following section 2.

2 Applications of emotional synthesis

In Figure 2 we display some possibilities of emotional processing in human machine interaction. This topic is further discussed in [7]. Emotions can be processed in several places of an information-processing system [8]:

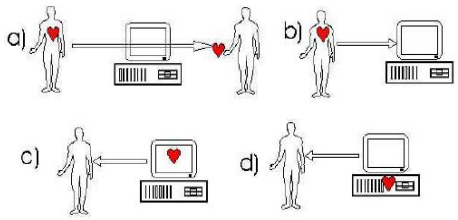

Figure 2 - Different uses of emotional processing in computer systems.

a) Broadcast: Emotional expression is an important channel of information in human communication. In telecommunication it might be desirable to provide for a special channel for emotional communication. A popular example are the so-called *emoticons* used in e-mail communication.

b) Recognition: The human emotional expression can be analyzed in different modalities, and this knowledge is used to alter the system reaction.

c) Simulation: Emotional expression can be mimicked by the system in order to enhance a natural interface or to access further channels of communication, like uttering urgent messages in an agitated speech style.

d) Modeling: Internal models of emotional representations can be used to represent user- or system states or as models for artificial intelligence, for example influence decision making.

In cases a), c) and d), emotional speech synthesis can be used to express or transmit emotional states. Thinking of applications, the following lists some ideas:

- fun, for example emotional greetings
- prosthesis
- emotional chat avatars
- gaming, believable characters
- adapted dialog design
- adapted persona design
- target-group specific advertising
- believable agents
- artificial humans

The list is ordered in an ascending time line we believe these applications will be realized. Because a technology has to be developed for a long time before it is stable and able to work under pressure, first applications will be about less serious topics like gaming and entertainment or will be adopted by users having a strong need, for example when used as prosthesis. In fact, these applications are already on the market, some links are given at
http://emotionalapplications.syntheticspeech.de/.

The applications further down the list are closely related to the development of artificial intelligence. Because emotions and intelligence are closely intermingled [9], great care is needed when computer systems appear to react emotional without the intelligence to meet the user's expectations with respect to dialog abilities.

Note that many of the use cases require the modeling of subtle speaking styles and addition of natural extra linguistic sounds to the synthesized speech. Considering the achievements in the past which mainly synthesized a set of exaggerated so-called basic emotions, there's still a long way to go.

3 Emotion modeling

The Emotional expression is usually either modeled by a categorical system, distinguishing between a specific set of emotion categories like *anger* or *boredom*, or by the use of emotional dimensions like *arousal*, *valence* or *dominance*. The categorical model has the charm of being intuitively to understand and being well established in human everyday communication.

With the dimensional model an emotion is modeled as a point in a multidimensional space describing emotional relevant dimensions like arousal or activation (ranging from relaxed to tense), pleasantness or valence (distinguishing between the subjective positiveness of an emotion) and dominance (which indicates the strength a person feels). Numerous additional dimensions have been suggested, but these three are traditionally the most common ones.

With speech synthesis in mind, it is easy to derive appropriate acoustic modification rules for the "arousal" dimension, because both are directly related to muscle tension, but very difficult for the other dimensions. Therefore emotional states are usually modeled by a categorical system, although dimensional systems are more flexible and better suited to model the impreciseness of the "real world", in which the so-called "full blown" emotions rarely occur.

Other models that are closer related to psychology and artificial intelligence, like appraisal theory or the OCC model, don't play a big role with respect to emotional speech synthesis today. This might change in the future when affective procecssing becomes more frequent in computer programs.

4 Annotation Formalisms

The problem of how to decide with which emotional attitude to speak a given text is out of scope of this paper. Nonetheless, a specific format is needed to determine the affect for a synthesizer. Current commercial synthesizers, for example by Loquendo, simply added emotional paralinguistic events to the unit-inventory, as described in an IBM article [10]. With the Loquendo TTS director, a tool to tune the text-to-speech system, expressive units can be selected from hierarchical drop-down menus.

A more complex approach was followed by the W3C incubator group for an emotional markup language [11]. This group aims at the development of a markup language for emotional expression usable for annotation, recognition, synthesis and modeling in human machine interaction. Here's an example:

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
 xml:lang="en-US">
  <voice gender="female">
    <prosody contour="(0\%,+20Hz)(10\%,+30\%)(40\%,+10Hz)">
      Hi, I am sad know but start getting angry...
    </prosody>
  </voice>
  <emotion>
    <category name="sadness set="basic" intensity="0.6"/>
    <timing start="10\%" end="50\%"/>
  </emotion>
  <emotion>
    <category name="anger" set="basic" intensity="0.4"/>
    <timing start="50\%" end="100\%"/>
  </emotion>
</speak>
```

Embedded in a SSML (speech synthesis markup language) tag, the synthesizer is instructed to change the vocal expression of the output from sadness to anger. The language is still under development but will hopefully provide a common basis for emotional research, enable easy data and sub component exchange as well as a distributed market for emotion processing systems.

Extensibility is possible, but the most frequently used models mentioned in section 3 like, categories, dimensions or appraisal models are already supported.

5 Perceptual clues in the speech signal

Features of the speech signal are spectral (the “sound” of the voice), prosodic (the “melody” of the speech), phonetic (kind of spoken phones, reductions and elaborations), ideolectal (choice of words) and semantic features.

In order to investigate the perceptual clues of emotional speech, we recorded a database with acted speech, the EmoDB [4], which is available for free download to conduct own experiments.

In Figure 3, spectrograms of several emotions, spoken by different actors but always the same sentence (*In sieben Stunden wird es soweit sein*), are displayed. A spectrogram shows the amplitude of frequencies over time. Although the Berlin database was recorded with 48 kHz, the frequency scale of these spectrograms is limited to eight kHz.

The displayed emotions diverge with respect to prosody, phonetic realization and voice quality, as can be seen from the different amplitudes in the frequency bands.

The modification of features like semantics and idiolect lie within the scope of the text generation system, whereas the acoustic features are within the control of the speech synthesizer itself. As will be shown in the next section, different synthesis approaches currently have a trade off between naturalness of the speech and flexibility with respect to speech manipulation.

6 Synthesis approaches

This section discusses the most important synthesis approaches in the current research landscape. The main difference results from a trade off with respect to unnatural but flexible parametric synthesis and natural sounding but inflexible (with respect to out-of-domain utterances) concatenation of audio samples.

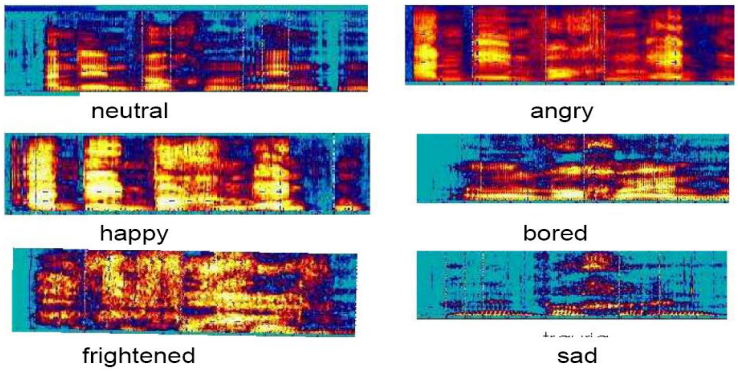

Figure 3 - Spectrograms of actors performing several emotions, always the same sentence (from [4]).

6.1 Rule based synthesis

In contrast to purely data-based statistical approaches, production models synthesize human speech by modeling aspects of the human production mechanism to a varying degree. These can be summarized by the term "system modeling" approaches, in contrast to "signal modeling" ones.

Essentially, articulatory models interpolate between target positions of the articulators while kinetic limitations with respect to velocity and degree of freedom are taken into account; the speech signal is then generated based on aerodynamic acoustic models. With the simulation of emotional speech in mind, articulatory synthesis is very attractive because the relation between the influence of emotional arousal on muscles and tissue and the acoustic properties of the body parts involved in the speech production process can be directly modeled.

However, it must be noted that articulatory synthesis is still suffering from an insufficiency of data. The correlation between speech and articulator movements is not well researched due to the lack of efficient measurement methods. Also the connection between laryngeal and articulator positions and the sound wave is highly complex; existing tube models are only a rough simplification of the human vocal tract.

Nevertheless, experiments with re-synthesis of natural speech and vowelconsonant- vowel logatoms show promising results [12], and clearly constrained phenomena like co-articulation can be studied in a controlled environment [13]. Furthermore, the data basis is improving due to advanced technology. For example, using new electromagnetic measurement methodologies, [14] have recently investigated the influence of emotional arousal on articulatory movement.

6.2 Unit selection synthesis

With the commercially most successful approach to speech synthesis, non-uniform unit selection, best fitting chunks of speech from large databases get concatenated, thereby minimizing a double cost-function: best fit to neighbor unit and best fit to target prosody. Because signal manipulation is reduced as much as possible, the resulting speech sounds most natural (similar to the original speaker) as long as the utterance to synthesize is close to the original domain of the database. In order to simulate emotional variation, three approaches are possible.

- Duplicate the database for each emotional style [15]. This can be seen as the "brute force" method and has been used successfully for prosthesis products, but of course only allows for the simulation of a very limited number of styles.

- Integrate an emotional target function in the unit selection process [5]. A very elegant method explicitly including the possibility to display extralinguistic speech sounds, but requires the manual annotation of very large databases.

- Apply signal manipulation methods on the speech signal [16]. This requires the units to be coded in a (semi-) parametric way. Voice transformation techniques can than be used to alter the speech style.

6.3 Synthesis based on statistical models

Statistical approaches combine the flexibility of the parametric synthesis with the naturalness of large databases. Currently, Hidden Markov Model (HMM)-based models are most successful [17]. The speech is modeled by a source-filter model and parametrized by an excitation signal and either Mel Frequency Cepstrum Coefficients, which are related to vocal tract shapes, or Line Spectrum Pairs (LSP), which are related to formant positions. The simulation of emotional styles is usually done by shifting the parameters of the source speech signal with respect to a target emotional style. With respect to the excitation signal, different voice qualities may be simulated by varying the parameters of a glottal flow model like the Liljencrants Fant model [18]. In Figure 4, several phonation types were simulated by formant synthesizer based on the LF-model [19].

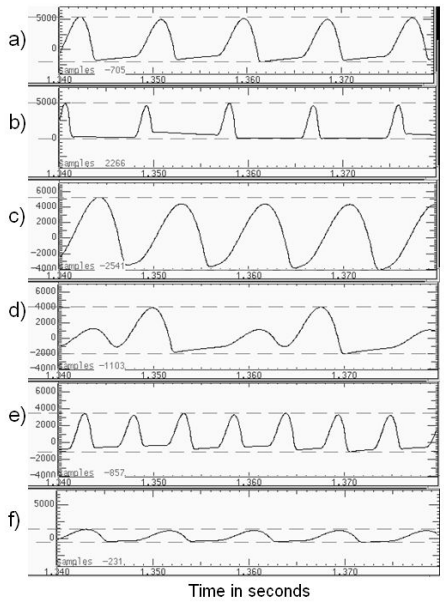

Figure 4 - Source signals for different phonation types. a: modal, b: tense, c: breathy, d: creaky, e: falsetto and f: whispery voice.

The greatest challenge in parametric synthesis is the lack of naturalness stemming from the vocoding stage of the process. Resynthesis of speech is possible without quality loss, but when either the source or the filter parameters are altered, audible distortions appear. The unnatural “buzziness” inherent in current parametric techniques may in part be due to insufficient models concerning source filter interaction.

6.4 Comparison of different approaches

The following table notes some peculiarities of different synthesis approaches with respect to emotional modeling.

	natural sounding	flexible modification	prosody manipulation	voice quality manipulation
formant synthesis	inherent buzziness due to inadequate source-filter modeling	very flexible due to full parametric control	rule based flexibility	rule based flexibility
articulatory synthesis	only copy synthesis possible	theoretically very flexible due to direct physical modeling of muscle tension	without limitation	yes
diphone-based synthesis	limited naturalness due to many concatenation points	only with respect to prosodic variation	yes, but high jumps introduce audible distortions	only by using several databases
non-uniform unit selection	quite natural within the domain of the database	no flexibility outside the scope of the database	only if emotional models are part of the concatenation function	only if emotional models are part of the concatenation function
HMM-based synthesis	introduces slight buzziness	yes, even linear interpolation between styles possible	yes, if emotional data is part of the database	yes, due to underlying source-filter model

7 Evaluation of emotional speech synthesizers

Emotional speech is getting evaluated with perception tests, usually given a forced choice test. Evaluation texts are usually emotionally neutral, although some authors criticize the inherent unnaturalness of this approach, and use emotional sentences [20]. The task for the judges would than be to rate the adequateness of the vocal expression, instead of a identification task.

As stated earlier, synthetic emotional expression can be exaggerated and results in high recognition rates of 80% and higher (of course depending on the number of emotions to identify), compared with emotion recognition.

Interestingly, the analysis of the resulting confusion matrices gives insights to the most prominent emotional dimensions in vocal emotional expression. Often, emotion pairs with a similar degree of activation like anger-joy or boredom-sadness get confused, whereas differences with respect to pleasantness and dominance seem to be primarily revealed in the facial expression.

8 Outlook

The challenge for the near future in this area is to improve both selection and conversion technology, and to make the unit selection approach more robust against missing data. In the very

long term, however, model-based approaches, which are inherently more flexible, may supersede unit selection technology if their quality approaches that of unit selection.
The main challenge for emotional speech synthesis results from the discrepancy between natural but inflexible vs. artificial sounding but flexible synthesis approaches. The solutions in the short to middle term consist of

- The usage of very large databases for non-uniform unit selection incorporating a cost function for emotional expression.
- The further development of statistically based hybrid parametric non-uniform unit selection techniques, specially with respect to naturalness.
- Promising are the advances in voice transformation techniques which can be used to introduce different speaking styles to speech resulting from a unit selection process.
- The development of high quality source filter model based synthesis like a wide band formant synthesizer or integrated source-filter modification models.

In order to reach the holy grail of speech synthesis: to have a synthesizer capable of modeling speaker characteristics from very small data, achievements in physical modeling techniques like articulatory synthesis are needed.
The two worlds: rule based physical modeling systems on the one hand and statistical algorithms based on very large data sets on the other, need to be combined to tackle demanding challenges like the simulation of affect in speech synthesis.

9 Acknowledgments

This work was partially funded by the EU FP6-IST project HUMAINE (Human-Machine Interaction Network on Emotion).

References

[1] I. R. Murray and J. L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” vol. 2, pp. 1097–1107, 1993.

[2] F. Burkhardt, *Simulation emotionaler Sprechweise mit Sprachsynthesesystemen*. Shaker, 2000.

[3] M. Schröder, “Emotional speech synthesis - a review,” in *Proc. Eurospeech 2001, Aalborg*, 2001, pp. 561–564.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Proceedings of Interspeech*, Lisbon, Portugal, 2005.

[5] N. Campbell, “Databases of expressive speech,” in *Proc. Oriental COCOSDA Workshop 2003, Singapore*, 2003.

[6] L. Devillers, R. Cowie, J.-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, “Real life emotions in french and english tv video clips: an integrated annotation protocol combining continuous and discrete approaches,” in *Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

[7] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth, "A taxonomy of applications that utilize emotional awareness," in *Proc. of the Fifth Slovenian and First International Language Technologies Conference Ljubljana*, 2006, pp. 246–250.

[8] R. Picard, *Affective computing*. MIT Press, 1997.

[9] A. R. Damasio, *Descartes' error: emotion, reason, and the human brain*. Avon Books, 1994.

[10] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to expressive speech synthesis," in *Proc. ISCA ITRW on Speech Synthesis, Pittsburgh*, 2003, pp. 79–84.

[11] M. Schröder, E. Zovato, H. Pirker, C. Peter, and F. Burkhardt, "W3c emotion incubator group report," http://www.w3.org/2005/Incubator/emotion/XGR-emotion/, 2008.

[12] P. Birkholz, D. Jackel, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," *Proc. ICASSP 2006 Toulouse*, 2006.

[13] P. Perrier, L. Ma, and Y. Payan, "Modeling the production of vcv sequences via the inversion of a biomechanical model of the tongue," in *Proc. Interspeech 2005 Lisbon*, 2005.

[14] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," *Proc. Interspeech 2005 Lisbon*, pp. 497–500, 2005.

[15] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "A speech synthesis system with emotion for assisting communication",," in *Proc. of the Isca Workshop on Speech and Emotion*, 2000, pp. 167–172.

[16] Y. Agiomyrgiannakis and O. Rosec, "Arx-lf-based source-filter methods for voice modification and transformation," in *Proc. of ICASSP*, 2009.

[17] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis." in *Proc. Eurospeech 1999, Budapest, Hungary.*, 1999, pp. 2347–2350.

[18] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," vol. 4, pp. 1–13, 1985.

[19] F. Burkhardt, "Rule-based voice quality variation with formant synthesis," in *Proc. Interspeech 2009, Bristol*, 2009.

[20] M. Schröder, *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, 2004.