Emotional Speech Synthesis

Felix Burkhardt and Nick Campbell

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 chapter discusses technical aspects of emotional speech synthesis, shows practical application and highlights new developments concerning the realization of affective speech with non-uniform unit selection based synthesis and voice transformation techniques.

Keywords/keyphrases

Speech synthesis, non-uniform unit-selection, hmm synthesis, voice transformation

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 perhaps lies in the complexity of human vocal expression: current state-of-the-art synthesizers still struggle with the challenge of generating understandable (for domain-independent systems) and natural sounding speech, although the latter requirement in itself already indicates the importance of affective expression.

This chapter gives 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. For a deeper understanding of the principles of speech synthesis see (Taylor 2009), for a deeper history of emotional speech synthesis, the reader is refered to (Murray & Arnott 1993) and (Schröder 2001).

The notion of emotional behavior is a most imprecise and difficult one to describe. As a psychologist once famously remarked: “everyone except a psychologist knows what an emotion is” (Young 1973, cited in Kleinginna & Kleinginna 1981).

Fortunately, this definition problem concerns emotion recognition much more than emotional speech synthesis, because in the former case one needs an accurate model of “the real world” in order to capture the multitude of emotional expressions, while in the latter, a simple model based on a few basic emotions might suffice for many applications, even if a “natural” expression rarely is of this form . Simulated emotional expression is typically very well recognized when limited to the display of some exaggerated prototypical emotions (Murray & Arnott 1993, Burkhardt 2000, Schröder 2001).

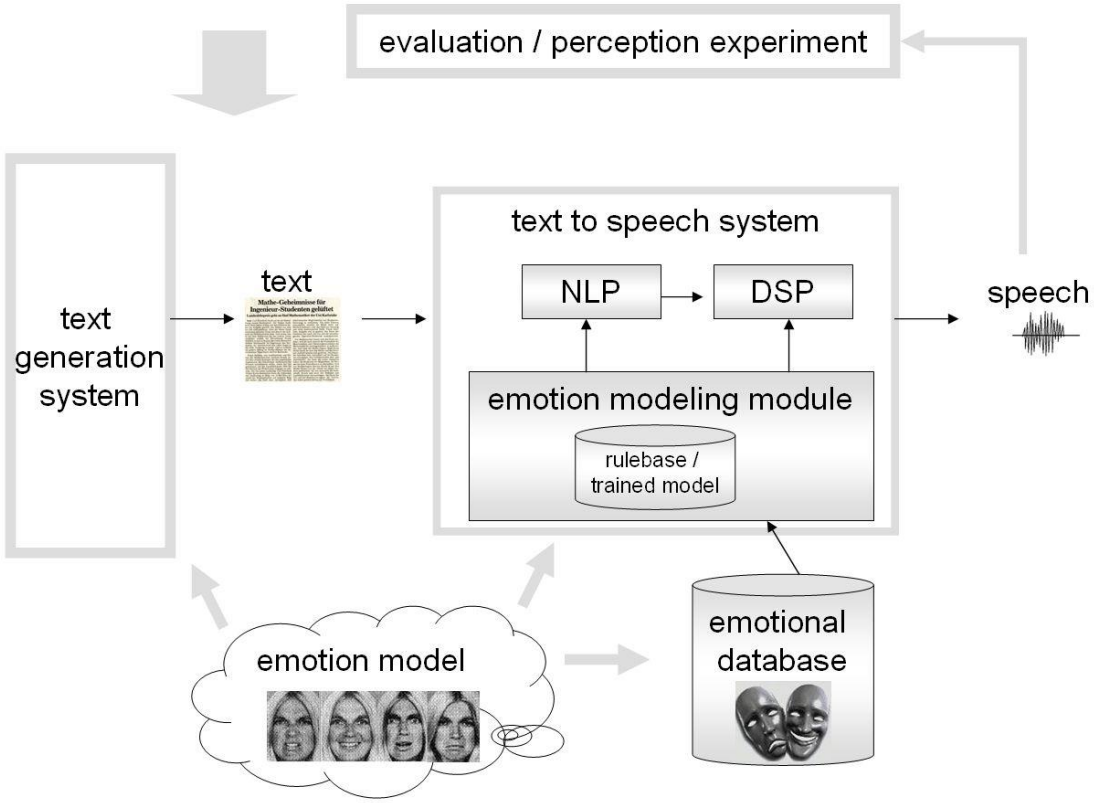

Figure 1: General architecture for an emotional text to speech synthesis system

Generally speaking, speech synthesis can be classified under the following three types (although hybrid forms are possible).

- Voice response systems as used in a recorded announcement of stops in public transport systems.,
- Re- or copy-synthesis as 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; these 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 beyond the scope of this article, formalisms to annotate text emotionally are discussed in Section 4 of this chapter as well as in chapter 29 on Emotion Markup Language.

A text-to-speech synthesizer converts text into speech by first analyzing the text by natural language processing (NLP) and conversion to 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 generating 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 (giving the meaning). All of these can be influenced by emotional expression, although, for practical reasons,

speech synthesis systems typically take only a subset of these features into account. Aspects of emotional expression affecting speech features are discussed in Section 5.

The component responsible for generating 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 are produced. Databases are often recorded by actors (Burkhardt, 2005), taken from real life data (Campbell, 2003), or from TV (Devilliers et al, 2006).

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.

A productive system must be evaluated by some means or other. 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, respectively the character appearance: 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 speech synthesis

Batliner et al discussed some ways of using emotional speech processing in (Batliner et al, 2006). Speech synthesis can be used to express or transmit emotional states which is important to a growing variety of use cases. The following lists some typical applications:

- 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 as we believe these applications will be realized. Because a technology has to be used for a long time before it is stable and able to work reliably under pressure, the early applications will include less serious domains of application like gaming and entertainment or will be adopted by users having a strong need, for example when used as prosthesis.

The applications further down the list are closely related to the development of artificial intelligence. Because emotions and intelligence are closely intermingled (Damasio, 1994), great care is needed when computer systems appear to react emotionally without having 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 of the past which mainly synthesized a set of exaggerated so-called basic emotions, there's still a long way to go.

3. Emotion modeling

As discussed in Chapter 2, the Emotional expression is usually either modeled by a categorical system, distinguishing between a specific set of emotion categories like anger, fear, or boredom, or by the use of emotional dimensions like arousal, valence or dominance. The categorical model has the advantage of being intuitively easy 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 emotionally relevant dimensions like arousal or activation (describing muscle relaxation and therefore being dircetly mappable to articulatory speech synthesizer parameters), pleasantness or valence (distinguishing between the subjective positivity of an emotion) and dominance (which indicates the strength of emotion a person feels). Numerous additional dimensions have been suggested, but these three are traditionally the most common ones (Rubin & Talerico 2009). With speech synthesis in mind, it is easy to derive appropriate acoustic modification rules for the "arousal" dimension, because this is 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 (Ortony et al, 1988), don't play a big role with respect to emotional speech synthesis today directly, although many appraisal theories posit basic emotions which corresponds to a categorical system (Chapter 5). For most of the use cases mentioned in the previous section, the simulation of "emotions" is perhaps not the most pressing requirement, but the ability of the synthesizer to express different "speaking styles"; these either derive from the use case and can be described by some "speaking style to acoustic property" rules, or are learned implicitly from data.

4. Annotation formalisms

The problem of how to decide with which emotional attitude to speak a given text is out of scope of this chapter. Nonetheless, a specific format is needed to

determine the display of affect for a synthesizer. For an overview article on this subject please see (Schröder et al, 2011).

Current commercial synthesizers, for example by Loquendo (now acquired by Nuance Inc.) or IBM, simply added emotional paralinguistic events to the unit-inventory, as described in an IBM article (Eide et al, 2003). 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 (Schröder et al, 2008, Chapter 29 in this book). This group aimed 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 now 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 an SSML (speech synthesis markup language, another standard proposed by the W3C) 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 above 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.

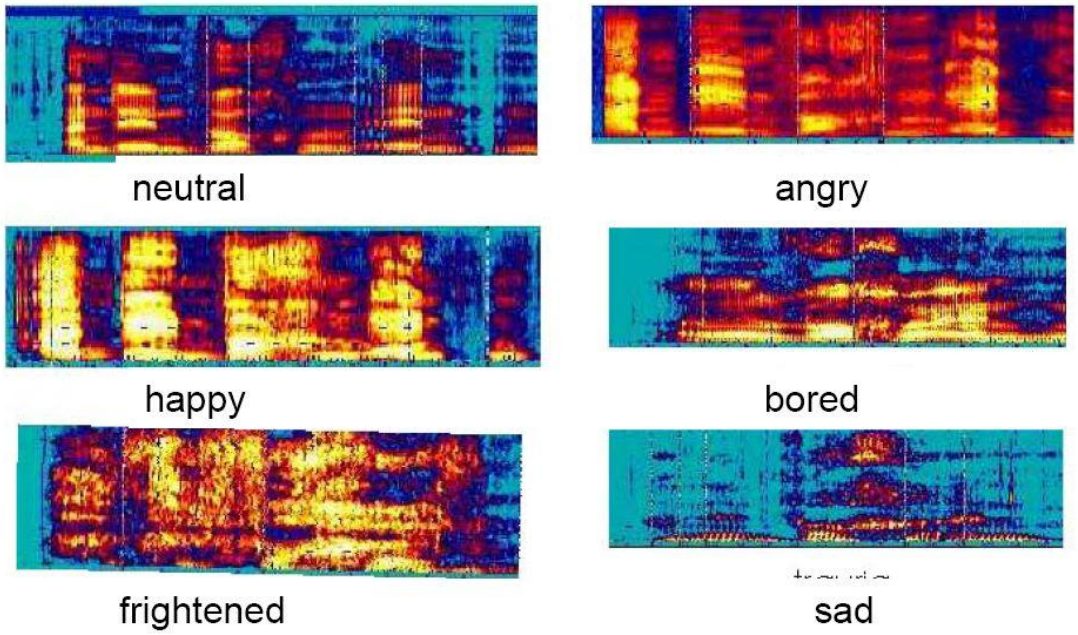

Figure 2: spectrograms from emotionally acted speech, always the same sentence

In order to investigate the perceptual clues of emotional speech, (Burkhardt et al, 2005) recorded a database with acted speech, generally known as the EmoDB, which is available for free download.

In Figure 2, spectrograms of several emotions, spoken by different actors but always the same sentence ("*In sieben Stunden wird es soweit sein*", "in seven hours it will happen"), 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 which is the normal upper limit of most synthesized speech.

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 common synthesis approaches in the current research landscape. The main difference between them can be characterised by 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. All of these are working in real time and could thus be used for human machine interaction. As explained in the following, the approaches differ in the possibility to simulate emotional arousal A general advice on which technology to use for emotional systems can not be given. The uncanny valley effect says that the more natural systems are not necessarily the ones that are accepted best by users. This means that more artificial sounding systems might be preferred in certain conditions.

Rule-based

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. Very high quality speech has been produced, but the processes are difficult to automate.

The correlation between speech and articulator movements is not well researched due to a lack of efficient data-capture and measurement methods. Also the connection between laryngeal and articulator positions and variations in the sound wave is highly complex; existing tube models represent only a crude simplification of the human vocal tract. Nevertheless, experiments with re-synthesis of natural speech and vowel-consonant- vowel logatoms show promising results (Birkholz et al, 2006), and clearly constrained phenomena like co-articulation can be studied in a controlled environment (Perrier et al, 2005). Furthermore, the data base is improving due to advanced technology. For example, using new electromagnetic measurement methodologies, (Lee et al, 2005) have recently investigated the influence of emotional arousal on articulatory movement.

Data-based

Diphone concatenation

Around 1986, the invention of the new signal processing technique PSOLA marked the birth of diphone synthesis, which proceeds by concatenating small units of recorded speech taken from a minimal set covering a given language. From the 90s on, diphone concatenation was the primary instrument for a large number of studies in emotional speech synthesis. Due to the nature of the decoding algorithms, only pitch and durational features could be modified, whereas voice quality was preserved by the natural waveform shaping. In order to control voice quality, there have been experiments with multiple diphone databases from the same speaker. Both formant and diphone synthesis used hand-crafted rules to modify the voice acoustics in order to express different emotions (Cahn, 1990 or Murray & Arnott, 1993)

Non-uniform unit-selection

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.

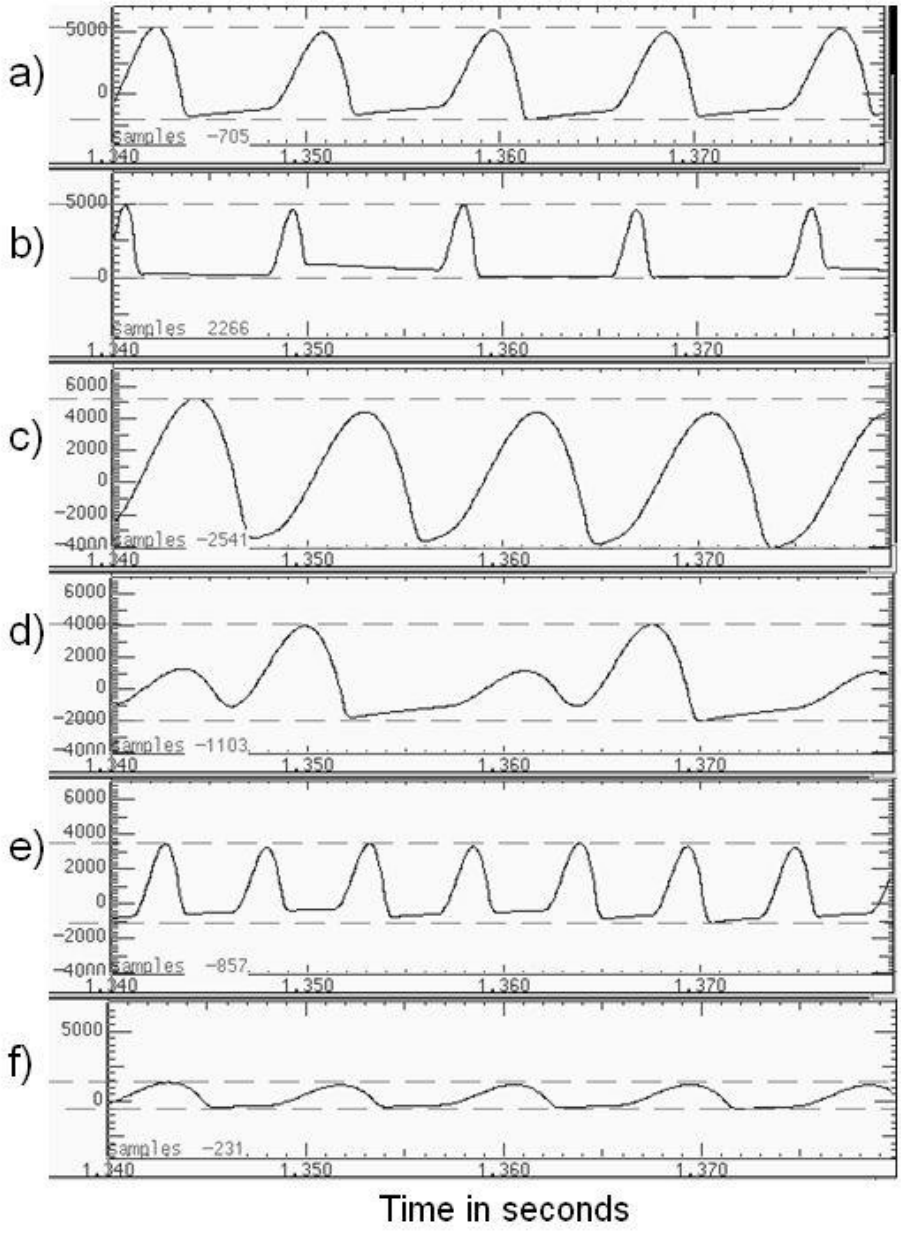

Figure 3: source signal waveforms for different phonation types. a: modal, b: tense, c: breathy, d: creaky, e: falsetto and f: whispery voice.

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 (Iida et al, 2000). 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 (Campbell, 2003). A very elegant method explicitly including the possibility to display extra linguistic speech sounds, but requires the manual annotation of very large databases.
- Apply signal manipulation methods on the speech signal (Agiomyrgiannakis & Rosec, 2009). This requires the units to be coded in a (semi-) parametric way. Voice transformation techniques can than be used to alter the speech style.

Hybrid approaches

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 (Yoshimura et al, 1999). The speech is modeled by a source-filter model and parameterized 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.

This approach inherits many of the tools and processes of automatic speech recognition, which models the speech sounds by means of a sequence of states, each representing part of a phoneme, with statistical probabilities learnt for the transitions between states and the mapping of these transitions onto sequences of words in a text. HMM-based synthesis reverses the process, to predict the sequence of sub-phonemic states from a given input text. It differs from standard speech recognition in that it includes a representation of phoneme duration and pitch as a characteristic of each state model and thus produces not just an acoustic sequence for synthesis but also an indication of the prosody of the utterance.

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 (Fant et al, 1985). In Figure 3, several phonation types were simulated by formant synthesizer based on the LF-model (Burkhardt, 2009).

Synthesis of non-verbal vocalizations

Beneath spoken words, human speech consists to a large degree of non-verbal vocalizations, which might be devided into “vegetative sounds”, “affect sounds”, “interjections” and “feedback and fillers sounds” (Trouvain & Truong, 2012). Since there is no standard to define and to classify (possible) non-speech sounds the annotations for these vocalizations differ very much for various corpora of conversational speech.

The analysis of these phenomena is a foundation to introduce non-verbal vocalizations into speech synthesis in order to make the output more natural. There seems to be agreement in that hesitation sounds and feedback vocalizations are considered as words (without a standard orthography) while the most frequent non-verbal vocalization are laughter on the one hand and, if considered a vocal sound, breathing noises on the other (Trouvain & Truong, 2012). Laughter and other feedback mechanisms are an essential part of normal face-to-face spoken interaction. These sounds are now beginning to be modeled both as part of the communication process and as components of advanced speech synthesis.

To date, except from some studies (Sundaram & Narayanan 2006), there has not been much effort devoted to reproducing the sounds of laughter in rule-based synthesis, but it is easy to include in concatenative methods if present in the source speech material (Eide et al, 2003). The classification of laughter remains as an active research field, but for conversational speech synthesis there will be a need for several different types of laughter, including embarrassed nervous laughs, gutsy humorous laughs, polite social laughs, and ice-breaking speech laughs. Ultimately, there may be a need to synthesise laughing speech as well, but that still remains as work for the future.

Feedback is essential in face-to-face communication, allowing the speaker to know that his or her utterance has been heard, processed, and understood, and this presents special challenges

for computer speech synthesis. Even the simple word "yes" can be said in many different ways, varying in meaning from a simple "go ahead" through the literal "I agree" even as far as "I hear you but completely DISagree" (when spoken with a slow rise-fall-rise intonation). The simplest sounds present the greatest problems for synthesis because their text does not indicate their meaning, and the intention of 'words' such as "um", "oh", "ah", etc. is carried largely by their prosody. There is a need for extension to the mark-up languages to allow indication of speaker (or usually listener) intention when specifying such sounds for synthesis.

Large source corpus-based emotional speech synthesis

Expressivity varies according to complex interactions of factors, with the speaker's emotional state being just one of many. In daily life, a person's voice and speaking style changes according to politeness, to their health, their various social and personal relationship(s) to the interlocutor, and to the context of the conversation.

Speech synthesis is reliant on good data to model these variations in order to reproduce the necessary tones and intonations required for interactive speech synthesis. Figure 3 shows how the voice changes during different speaking styles, and this range of variation is common in everyday speech. The types of recordings made for early speech synthesis used actors and were recorded in a studio, often from textual prompts. More recent recordings are capturing the speech of ordinary people in a wide range of everyday situations. The recording technology is now cheap and ubiquitous, but the manual effort required for the annotation of such recordings is still very expensive. That said, the amount of effort required in producing reliable annotations is finite and can result in a rich source of material for resynthesis. Furthermore, by massively increasing the amounts of recorded speech, we then find multiple tokens of the same phonemic sequence but with subtle differences in intonation , prosody, and voice quality, such that they become ideal for use in rich expressive synthesis without the need for complex signal modifications. The challenge, as said above, is to label them

appropriately so that suitable tokens can be retrieved form the corpus to express the target intentions and cognitive states of the speaker.

Technology might come to our aid here as the dependencies between cognitive state and speech acoustics become known and as the ability to automatically recognise and label different voice qualities and prosodic contours improves. Currently, little use is made of the natural variety in voice characteristics in a large source corpus; but by sophisticated selection and careful concatenation, we may find that recordings of everyday speech provide all the necessary material for complex expression of emotion and interlocutor relationships. Rather than go for larger corpora, it may be wiser to make finer use of the variability we already have.

Evaluation

Emotional speech is usually evaluated with perception tests, often of the forced choice variety.

Evaluation texts are designed to be emotionally neutral, although some authors criticize the inherent unnaturalness of this approach, and prefer to use emotional sentences (Schröder 2004).

The task for the judges then is to rate the adequateness of the vocal expression, instead of a simple (categorical) identification task.

As stated earlier, synthetic emotional expression tends to be exaggerated and this 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.

As the demands of speech synthesis go beyond cartoon-voice generation and it finds uses in everyday conversational situations, these extreme stereotypical emotions will be less welcome and the need will be to synthesize subtle differences in voice and prosody that signal more complex speaker-hearer interactions. For this work, simple forced-choice evaluations will give way to more sophisticated measures of appropriateness for a given context or situation and more subtle statistical processes to provide diagnostic as well as likeability information.

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, possibly based on voice quality and prosodic characteristics as selection criteria.
- The further development of statistically based hybrid parametric non-uniform unit selection techniques, especially 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.
- The ability to automatically categorise voice characteristics and to include these alongside phonemic features in the training or selection of material.

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. until then, the manual labour required in the careful annotation of large recordings of natural

conversation al speech continue to produce high-quality output and competitive evaluations like the Blizzard Challenge confirm that the winners typically are those who devote more time to careful annotation of the data and manual cleaning-up of the training material before starting with the more technical aspects of synthesiser production.

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.

References

Agiomyrgiannakis, Y., & Rosec, O. (2009). Arx-lf-based source-filter methods for voice modification and transformation, in Proc. of ICASSP.

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

Birkholz, P., Jackel, D., & Kröger, B. J. (2006). Construction and control of a three-dimensional vocal tract model, Proc. ICASSP, Toulouse.

Burkhardt, F. (2000). Simulation emotionaler Sprechweise mit Sprachsynthesesystemen. Shaker.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of german emotional speech, in Proceedings of Interspeech, Lisbon, Portugal.

Burkhardt, F. (2009). Rule-based voice quality variation with formant synthesis, in Proc. Interspeech, Brighton.

Cahn, J. E. (1990). The Generation of Affect in Synthesized Speech. Journal of the American Voice I/O Society, 8:1–19.

Campbell, N. (2003). Databases of expressive speech, in Proc. Oriental COCOSDA Workshop, Singapore.

Damasio, A. R. (1994). Descartes' error: emotion, reason, and the human brain. Avon Books.

Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, S., & McRorie, M. (2006). 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.

Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., & Pitrelli, J. (2003). A corpus-based approach to expressive speech synthesis, in Proc. ISCA ITRW on Speech Synthesis, Pittsburgh, pp. 79–84.

Fant, G., Liljencrants, J., & Lin, Q. (1985).A four-parameter model of glottal flow, vol. 4, pp. 1–13.

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

Kleinginna P. R., & Kleinginna, A. M. (1981). A Categorized List of Emotion Definitionns, with Suggestions for a Consensual Definition. Motivation & Emotion, pp. 345–379.

Lee, S. , Yildirim, S., Kazemzadeh, A., & Narayanan, S. (2005). An articulatory study of emotional speech production, Proc. Interspeech 2005 Lisbon, pp. 497–500.

Murray, I. R.,& Arnott, J. L.(1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, vol. 2, pp. 1097–1107.

Ortony, A., Clore, G. L., and Collins, A. (1988). The Cognitive Structure of Emotion. Cambridge University Press, Cambridge, UK.

Perrier, P., Ma, L., & Payan, Y. (2005). Modeling the production of vcv sequences via the inversion of a biomechanical model of the tongue, in Proc. Interspeech, Lisbon.

Rubin, D. C., & Talerico, J.M. (2009). "A comparison of dimensional models of emotion". Memory 17: 802–808.

Schröder, M. (2001). Emotional speech synthesis - a review, in Proc. Eurospeech, Aalborg, pp. 561–564.

Schröder, M. (2004). 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.

Schröder, M., Zovato, E., Pirker, H., Peter, C., & Burkhardt, F. (2008). W3c emotion incubator group report, http://www.w3.org/2005/Incubator/emotion/XGR-emotion/.

Schröder, M., Pirker, H., Lamolle, M., Burkhardt, F., Peter, C., & Zovato, E. (2011). Representing emotions and related states in technological systems. In P. Petta, R. Cowie, & C. Pelachaud (Eds.), Emotion-Oriented Systems -- The Humaine Handbook (pp. 367-386). Springer.

Sundaram, S., & Narayanan, S. (2006). Automatic acoustic synthesis of human-like laughter, JASA, Pages: 527–535.

Taylor, P. (2009). Text to Speech synthesis, Cambridge University Press.

Trouvain, J. & Truong, K. 2012. Comparing non-verbal vocalisations in conversational speech corpora. Proc. 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals, Istanbul, pp. 36-39.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. in Proc. Eurospeech, Budapest, Hungary,, pp. 2347–2350.