Expressive Speech Synthesis: Past, Present, and Possible Futures

Marc Schröder

DFKI GmbH, Saarbrücken, Germany
http://dfki.de/~schroed

Abstract. Approaches towards adding expressivity to synthetic speech have changed considerably over the last 20 years. Early systems, including formant and diphone systems, have been focussed around "explicit control" models; early unit selection systems have adopted a "playback" approach. Currently, various approaches are being pursued to increase the flexibility in expression while maintaining the quality of state-of-the-art systems, among them a new "implicit control" paradigm in statistical parametric speech synthesis, which provides control over expressivity by combining and interpolating between statistical models trained on different expressive databases. The present chapter provides an overview of the past and present approaches, and ventures a look into possible future developments.

1 Introduction

Synthetic speech nowadays is mostly intelligibly; in some cases, it also sounds so natural that it is difficult to distinguish it from human speech. However, the flexible and appropriate rendering of expressivity in a synthetic voice is still out of reach: making a voice sound happy or subdued, friendly or empathic, authoritative or uncertain are beyond what can be done today. Indeed, even though quite some research has been carried out on the issue of adding any kind of recognisable expressivity to synthetic speech, the question of its "appropriateness" in a given communicative situation has barely been touched upon.

This chapter aims to give an overview of the work in the area in the past and in present work, and points to some types of developments that may bring sustantial contributions to the area in the future.

Adding expressivity to a synthetic voice has two aspects: being able to realise an expressive effect, i.e. having the technological means to control the acoustic realisation; and having a model of how the system's realisation should sound in order to convey the intended expression. We will see that some progress has been made on both of these levels, but that neither the technological means of control nor the models of expressivity in state-of-the-art systems are anywhere close to what would be needed for truly natural expressive speech synthesis. Nevertheless, it is interesting to look at existing approaches to understand the challenges involved.

The chapter is structured in terms of technologies as they evolve over time. Through that structure, it will become clear that the technologies are not independent of the approach chosen for modelling expressivity, but that each technology has its favourite and natural control model – "explicit control" and "playback" in early times, and "implicit control" in some of the current works.

2 A short history of expressive speech synthesis

Work on emotional expressivity in synthetic speech goes back to the end of the 1980s. As technology evolved, different technologies have been used, each with their strenghts and weaknesses. The following sections provide a concise overview of the range of techniques employed over time; for more details, see [1].

2.1 Formant synthesis

Rule-based formant synthesis, the oldest speech synthesis technique, predicts the acoustic realisation of speech directly using a set of rules, usually designed carefully by experts. As these rules tend not to capture the complexity of natural human speech, the synthetic speech sounds unnatural and "robot-like". By its very nature, however, formant synthesis allows for the control over a broad range of parameters, including glottal as well as supraglottal aspects of speech production. Many of these parameters are potentially relevant for modelling expressive speech.

The first emotionally expressive speech synthesis systems were created based on the commercial formant synthesiser DECTalk: both Cahn's Affect Editor [2] and Murray et al.'s HAMLET [3] added emotion-specific acoustic modification modules on top of that system. For each emotion category, they implemented an explicit model for the acoustic realisation, drawing on the existing literature and fine-tuning the rules in a trial-and-error procedure.

In a more recent study, Burkhardt [4] used formant synthesis to systematically vary acoustic settings in order to find perceptually optimal values for a number of emotion categories. A first, coarse exploration of a five-dimensional acoustic space was followed by an optimisation phase adding parameters based on the literature in order to increase the specificity of emotion patterns.

Examples for the kinds of explicit emotion models implemented in these systems are listed in Table 1.

2.2 Diphone concatenation

In the early 1990s, a fundamentally different synthesis technology gained popularity: synthesis by concatenation, i.e. by re-sequencing small pieces of human speech recordings. The early systems recorded one example of each "diphone" – the piece of a speech signal going from the middle of one phone to the middle of the next phone –, usually at monotone pitch. By re-ordering these, any

phone sequence in a given language can be generated. Through signal processing techniques such as pitch-synchronous overlap-add (PSOLA) [5], different F0 contours and duration patterns could be realised at the price of a certain degree of distortion. The resulting quality is usually quite a bit better than formant synthesis; however, it is important to note that the voice quality (tense, creaky, lax, etc.) is fixed, determined by the rendition of the speaker during diphone recordings.

Therefore, when using diphone synthesis for expressive speech, it is a crucial question whether it is acceptable not to modify voice quality, i.e. whether F0 and duration alone can create a desired expressive effect. Interestingly, different studies have come to different conclusions on this matter. A majority (e.g., [6], [7], [8]) report at least some degree of success in conveying various emotions; some, however, report recognition rates close to chance level ([9], [10]). Indeed, by cross-combining voice quality and prosody of different expressive styles, Montero et al. [8] and Audibert et al. [11] found that for a given speaker, the relative contribution of prosody and voice quality depends on the emotion expressed. However, there is not yet a clear picture which emotions predominantly rely on voice quality and which can be recognised based on prosody. Indeed, it may be [12] that different speaker strategies exist to express the same emotion, relying to various extents on voice quality and on prosody.

In the absence of explicit parametric control over voice quality, one approach is to record the diphone inventory with the same speaker's voice in different voice qualities [13]; in order to change the voice quality, one needs to switch to a different diphone inventory from the same speaker.

2.3 Explicit prosody control

The synthesis methods described above require explicit settings for prosodic parameters in order to convey emotional expressivity. How are these settings determined?

A straightforward approach to control the acoustic correlates of a given emotion is to copy a natural rendition, the so-called "copy synthesis" (e.g., [14]): the parameters that can be controlled in the synthesizer (e.g., F0 and duration) are measured in an expressive recording, and used directly as the input to a synthesizer. This method is appropriate for testing to which extent a given synthesis method can reproduce the same auditory impression as a human speech sample; however, the generalisability is obviously low.

A general text-to-speech setting requires an automatic way of computing the acoustic correlates of an expression from a representation of an emotion. With formant and diphone synthesis, such a model needs to link explicitly the emotion with its acoustics, in the sense that the parameters through which the emotion is expressed are clearly listed and the respective effect is explicitly stated. As we will see below, this contrasts with the implicit modelling introduced in recent statistical synthesis approaches.

Explicit prosody models have been formulated based on various sources of information. Usually, rules are based on effects reported in the literature (e.g.,

[2]), on analyses of own data (e.g., [15]), or on perception studies (e.g., [4]), and are fine-tuned in a manual trial-and-error procedure. Clearly, such rules can only be as good as the quality of the literature reports, data analysed, and the analysis and control parameters studied.

Table 1 lists some examples of explicit prosody rule sets as described in the literature, for a typical selection of emotional states modelled (from [16]).

Systems vary with respect to the underlying representation of the "predictor" variables, e.g. emotions: while most studies used emotion categories or other distinct states, it is also possible to use the emotion *dimensions* arousal, valence and power [18].

It is worth pointing out that even the relevant set of acoustic parameters is a research issue, and is far from being resolved. All studies in the area seem to agree on the importance of global prosodic settings, such as F0 level and range, speech tempo and possibly loudness. Some studies try to go into more detail about these global settings, modelling e.g. steepness of the F0 contour during rises and falls, distinguishing between articulation rate and the number and duration of pauses, or modelling additional phenomena like voice quality or articulatory precision. A further step is the consideration of interactions with linguistic categories, like further distinguishing between the speech tempo of vowels and consonants, of stressed and unstressed syllables, or the placement of pauses within utterances. The influence of linguistic prosodic categories, like F0 contours, is only rarely taken into account, although these have been shown to play an important role in emotion recognition [19, 4].

2.4 Unit selection

In the mid-1990s, the concept of concatenative synthesis was developed further: instead of using just one example of each diphone recorded with a flat pitch, several versions of a diphone *unit* are recorded in natural speech. Through a sophisticated selection method, the most suitable chain of such units is determined for any given target sentence – hence the name *unit selection* synthesis. If suitable units are available in the recordings, no or very little signal processing is needed, and the resulting synthesized speech can sound highly natural, sometimes difficult to distinguish from natural human speech.

For expressive speech, the strength of the method – interfering with the recorded speech as little as possible – is also its main weakness: only the speaking style recorded in the database can be generated, and little or no control is available over prosody and voice quality.

Nevertheless, it is of course possible to generate expressive speech within the unit selection framework, in what could be called a "playback" approach: in order to generate speech in style X, a database in style X is recorded. In this way, Iida and Campbell [20] have produced a system capable of speaking with the same person's voice in each of three emotions: happy, angry, and sad. To achieve the effect, the authors had to record three databases: one consisting of utterances spoken in a happy tone, one spoken in an angry tone and one spoken

Emotion Study Language Rec. Rate	**Parameter settings**
Joy Burkhardt and Sendlmeier [4] German 81% (1/9)	**F0 mean**: +50% **F0 range**: +100% **Tempo**: +30% **Voice Qu.**: modal or tense; "lip-spreading feature": F1 / F2 +10% **Other**: "wave pitch contour model": main stressed syllables are raised (+100%), syllables in between are lowered (-20%)
Sadness Cahn [2] American English 91% (1/6)	**F0 mean**: "0", reference line "-1", less final lowering "-5" **F0 range**: "-5", steeper accent shape "+6" **Tempo**: "-10", more fluent pauses "+5", hesitation pauses "+10" **Loudness**: "-5" **Voice Qu.**: breathiness "+10", brilliance "-9" **Other**: stress frequency "+1", precision of articulation "-5"
Anger Murray and Arnott [3] British English	**F0 mean**: +10 Hz **F0 range**: +9 s.t. **Tempo**: +30 wpm **Loudness**: +6 dB **Voice Qu.**: laryngealisation +78%; F4 frequency -175 Hz **Other**: increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear Burkhardt and Sendlmeier [4] German 52% (1/9)	**F0 mean**: "+150%" **F0 range**: "+20%" **Tempo**: "+30%" **Voice Qu.**: falsetto
Surprise Cahn [2] American English 44% (1/6)	**F0 mean**: "0", reference line "-8" **F0 range**: "+8", steeply rising contour slope "+10", steeper accent shape "+5" **Tempo**: "+4", less fluent pauses "-5", hesitation pauses "-10" **Loudness**: "+5" **Voice Qu.**: brilliance "-3"
Boredom Mozziconacci [17] Dutch 94% (1/7)	**F0 mean**: end frequency 65 Hz (male speech) **F0 range**: excursion size 4 s.t. **Tempo**: duration rel. to neutrality: 150% **Other**: final intonation pattern 3C, avoid final patterns 5&A and 12

Table 1. Examples of successful explicit prosody rules for emotion expression in synthetic speech (from [16]). Recognition rates are presented with chance level for comparison. Sadness and Surprise: Cahn uses parameter scales from –10 to +10, 0 being neutral; Boredom: Mozziconacci indicates intonation patterns according to a Dutch grammar of intonation, see [17] for details.

in a sad tone. A given emotion was generated by selecting units only from the corresponding subset of the recordings.

Johnson et al. [21] pursued a similar approach. They employed limited domain synthesis for the generation of convincing expressive military speech, in the framework of the Mission Rehearsal Exercise project. The styles, each recorded as an individual limited domain speech database, were shouted commands, shouted conversation, spoken commands, and spoken conversation.

Along the same line of thought, Pitrelli et al. [22] recorded full unit selection databases for a "good news" vs. a "bad news" expressive voice by the same speaker.

3 Current trends in expressive speech synthesis

It becomes clear from this short historic overview of formant, diphone, and unit selection synthesis that these technologies form a dichotomy, between flexibly parameterisable systems on the one hand, requiring explicit acoustic models of expressivity which are difficult to formulate, and natural-sounding but inflexible "playback" systems on the other hand.

In recent times, a range of approaches are being pursued to overcome this separation, by adding more control to data-driven synthesis methods in various ways. This section reviews some of the major trends.

3.1 Expressivity-based selection of units

In unit selection synthesis, the notion of a "target" is central – units that are similar to the given target are likely to be selected. Originally, this target is defined in purely linguistic terms: phone identity, context, position in the sentence, etc. Some systems also employ acoustic targets, in the sense of prosodic models trained on the synthesis database.

One important step towards flexible expressivity in unit selection synthesis would therefore be the use of expressivity-related targets in the selection process itself. Such targets can be of two kinds: symbolic and acoustic. A symbolic target can be a label identifying the intended speaking style; strictly enforced, it amounts to the same kind of separation between expressive subsets as used in earlier works. However, it can be complemented with a cost matrix, which for a given target style allows for the selection of a unit with a different style at a cost. For example, it would be possible to allow both "angry" and "neutral" units for a "sad" target, but to discourage the use of "angry" units by giving them a high penalty. This approach was used in recent work by Fernandez and Ramabhadran [23].

Acoustic expressive targets, on the other hand, rely on acoustical models of expressive styles to identify units that could be suitable for the targeted expressive style. A major attraction of this approach is that it could be used with unlabelled or only partially labelled databases.

The first attempt in this sense seems to have been published by Campbell and Marumoto [24]. They used parameters related to voice quality and prosody as emotion-specific selection criteria. Three emotion-specific databases [20] were combined into one database. Different kinds of selection criteria were tested, including Hidden Markov Models (HMMs) trained on the emotion-specific subsets, and hand-crafted prosody rules. Recognition results in listening tests indicated a partial success: anger and sadness were recognised with up to 60% accuracy, while joy was not recognised above chance level.

Fernandez and Ramabhadran [23] explored a semi-automatic recognition of *emphasis* for synthesis. A statistical acoustic model of emphasis was trained on a manually labelled database consisting of approximately 1,000 sentences; this model was then used to identify emphasized words in a larger, unlabelled database consisting of about 10,000 sentences. Different symbolic labels were assigned for manually vs. automatically labelled emphasis; in a cost matrix, for a target style "emphasis", manually labelled emphasis units are slightly favoured over automatically labelled ones. In a listening test, the authors could show that the automatic annotation in the large corpus increased the number of sentences where emphasis was perceived as intented, even if the total level stayed very low (at 51% in a two-class setup).

In the context of generating expressive synthetic speech from the recordings of an audio book, Wang et al. [25] have prepared the ground for future work by identifying a suitable perceptual annotation scheme for the speech material. Out of the 18,000 utterances of the audio book, a representative sample of 800 expressive utterances was annotated using auditory ratings of pitch, vocal effort, voice age, loudness, speaking rate, and speaking manner. In addition, similarity ratings of utterance pairs were obtained. A set of nine perceptually homogeneous speaking styles was determined by growing a classification tree, using similarity ratings between sentences as an impurity criterion and auditory ratings as decision nodes. The next step in their work is to train acoustic classifiers on the nine utterance clusters, and to use those classifiers to annotate the full database, which would allow for expressive unit selection in a way similar to Fernandez and Ramabhadran.

3.2 Unit selection and signal modification

Signal modification, traditionally used in diphone synthesis, is usually avoided in unit selection because of the deteriorating effect on the overall speech quality. Nevertheless, it is possible to apply the same techniques, such as PSOLA, on unit selection output, modifying pitch and duration according to emotional prosody rules [26]. As for diphone synthesis, this approach has the disadvantage of not being able to modify voice quality, and of creating audible distortions for larger modifications.

Research is underway to improve on that state of things. Working towards an explicit control of the voice source, d'Alessandro and Doval [27] have proposed a method for modifying the glottal source spectrum, described by the parameters glottal formant and spectral tilt. They decompose the speech signal

into a periodic and an aperiodic part, and recombine these after modifying them separately.

One of the main problems in model-based modification of voice quality is the reliable automatic estimation of voice source parameters from the speech signal. Avoiding the problem of inverse filtering, Vincent et al. [28] have proposed an analysis by synthesis method for the joint estimation of glottal source and vocal tract filter coefficients: for each speech frame, they try out all examples of a codebook of glottal source configurations, and use the one with which a linear predictive coding (LPC) analysis yields the smallest residual error. The method has been applied to the analysis of emotional speech [11] and could potentially be used for signal modification.

A data-driven alternative to such explicit modelling attempts is to use voice conversion techniques with expressive speech. The main difference is that in voice conversion, a source and a target need to be defined by data sets. The aim is to define a mapping function that can convert a large, hiqh quality source (e.g., a large neutral unit selection voice) into a target style for which only a small amount of recordings would be required. Carrying out research in this direction, Matsui and Kawahara [29] converted one emotional recording of a speaker into a different emotion uttered by the same speaker. They achieved good recognition rates, but anchor points in the time-frequency plane had to be set manually. Ye and Young [30] used voice morphing for converting one speaker's voice into another speaker's, by means of linear transformations in a sinusoidal model. The same technology could be used to convert a given speaker's voice expressing one emotion into the same speaker's voice expressing another emotion.

A further step towards flexible control in the data-driven framework is the use of interpolation technology – the creation of a mixed speech signal out of two different speech streams. A method for interpolating the spectral envelope of a source and a target voice was proposed by Turk et al. [31]. They used diphone voices with different vocal effort [13], and created intermediate voices. In a perception test, these were rated as intermediate in vocal effort to the respective source and target voices. The method can be incorporated into unit selection [32], allowing for the continuous interpolation between a source and a target synthesis voice. The principle of interpolation is the same whether the target voice is recorded as in the "playback" approaches described above, or whether it is the result of a mapping process in a voice conversion framework.

3.3 HMM-based parametric speech synthesis

A new synthesis technology is establishing itself as a serious alternative to unit selection synthesis: statistical parametric synthesis based on Hidden Markov Models (HMMs). The method became very well known when it became the surprise winner of the first Blizzard speech synthesis competition in 2005 [33].

In essence, the method works as follows. Context-dependent HMMs are trained on a speech database; the spectrum, F0 and duration are modelled separately. The context-dependent models are organised in a decision tree; at run-time, for a given "target" context to be realised, the tree yields the appropriate HMM state

sequence corresponding to that context, describing mean and standard deviation of the acoustic features. A vocoding technique is used to generate an audio signal from the acoustic features, resulting in a very intelligible, but muffled-sounding speech output.

By nature, the approach is an interesting hybrid between data-driven and parametrisable synthesis approaches: similarly to unit selection, the approach is relatively natural-sounding because it is trained on natural speech data, but unlike unit selection, the speech is generated from a parametric representation which can potentially be controlled if a suitable "handle" to the parameters is provided.

As in unit selection, the simplest approach to expressive speech with HMMs is to train models on speech produced with different speaking styles. Yamagishi et al. [34] trained HMM-based voices from four different speaking styles: "reading", "rough", "joyful", and "sad". In addition to creating fully separate voices for each style, they also built a combined voice in which the style was part of the context description. For both kinds of voices, very high recognition rates were obtained in a perception test, showing that the basic technology can reproduce the style.

While in theory it would be possible to influence the parametric representation generated by the HMMs in the sense of explicit models similarly to formant and diphone synthesis, work on flexible style control with HMM-based synthesis has pursued a different method. It could be called an "implicit modelling approach", because landmarks for interpolation are defined from data rather than by explicitly characterising them as acoustic configurations. In this line of thought, Miyanaga et al. [35] presented a method to control speaking style by training style-specific HMM models and interpolating between them using a "style vector". Experiments were carried out that demonstrated, also for intermediate style values, a gradual effect on perceived style, for models trained on reading, joyful, rough, and sad speaking styles.

A further strength of HMM-based synthesis is that speaker-specific or style-specific voices can also be created by *adaptation* rather than training. While in training, a style-specific voice requires several hundreds of sentences spoken in a given speaking style, the adaptation of an average voice to a specific style can be performed with as little as a few dozen sentences. Yamagishi et al. [36] compared various adaptation techniques both in objective terms (objective distance to the target speech) and in terms of subjective similarity. Interestingly, they found that when creating a voice by *adapting* spectrum, F0 and duration of an average voice to 100 sentences from a target speaker, it sounded more similar to that target speaker than a voice *trained* on 453 sentences from that speaker. This is probably due to the greater robustness of the average voice model, trained on several thousands of sentences from various speakers. Note that from the point of view of technology, the terms "speaker" and "style" can be used interchangeably: if it is possible to adapt an average voice to a range of different speakers, it is very likely that it is also possible to adapt it to various speaking styles of one speaker.

In summary, even though this approach requires style-specific recordings as in unit selection, the resulting flexibility is greater. Because of the parametric representation, interpolation between the styles is possible, enabling the expression of low-intensity emotions by interpolation between a neutral speaking style and a high-intensity emotional speaking style. Similarly, blends of emotions can be generated by interpolating between two emotional speaking styles.

Table 2 summarises the approaches to parameterisation of expressive speech as found in current research on speech synthesis.

	Unit selection		**HMM-based synthesis**
	Selection of units	**Signal modification**	
explicit acoustic models	hand-crafted prosody rules as targets [24]	PSOLA with explicit rules [26] explicit control of glottal spectrum [27] brute-force estimation of glottal source parameters [28] (potentially usable for modification)	
playback	recording separate expressive voice databases [20, 21] manually labelled symbolic targets [23]		train models on style-specific speech data [34] adapt models to style-specific speech data [36]
implicit acoustic models	HMM models as targets [24] automatically trained symbolic targets [23], potentially [25]	voice conversion between emotional recordings of same speaker [29], potentially [30] interpolation between recorded or converted voices [31, 32]	interpolate between style-specific models [35]

Table 2. Current approaches to the parameterisation of expressivity in state-of-the-art speech synthesis

3.4 Non-verbal vocalisations

As research on expressive speech is moving towards conversational speech, it is becoming clear that read speech is an insufficient basis. Expressive conversa-

tional speech does not result from applying suitable prosody modifications to read speech. Indeed, as Campbell [37] showed, a large proportion of everyday vocalisations are non-verbal – laughs, speech "grunts", and other small sounds which are highly communicative but which are not adequately described as "text plus prosody".

To illustrate the point, Campbell [38] has produced a conversational speech synthesizer which uses a huge database of everyday speech as the unit selection database. With adequate annotation of speech units and careful manual selection, this system can produce conversational speech of unprecedented naturalness. However, the subtleties of meaning produced and perceived by humans with respect to the meaning of non-verbal vocalisations in context are still far beyond the reach of current automatic selection algorithms.

The difficulty of identifying "suitable" non-verbal exprssions for the use with synthetic speech was illustrated by Trouvain and Schröder [39]. They inserted laughter sounds of different intensity into a short dialogue, and asked listeners to rate their appropriateness. The result clearly showed that the lowest-intensity laughs were perceived as most suitable. In a similar line of research, Schröder et al. [40] investigated the suitability of various affect bursts (short emotional interjections, see [41] for details) as listener feedback in a short natural language interaction. Suitability ratings were interpreted in terms of social appropriateness, of "display rules" [42], to be obeyed by a listener in a conversation.

From these works, it becomes clear that research on the use of non-verbal vocalisations in human-machine communication is still in its infancy, but further work will be inevitable if truly natural conversational speech synthesis is the goal.

4 Possible future directions for expressive speech synthesis

So where is research on expressive speech synthesis heading? Will there be methods developed that are able to combine high quality with flexible control? Will they provide explicit or implicit control, in the sense of the words introduced above? While of course it is impossible to predict the future, there are some tendencies visible or emerging today, which it is worth having in mind when thinking about future developments.

One of the promising extensions of today's research is the improvement of vocoding algorithms for HMM-based speech synthesis. With more natural speech generation methods, parametric systems would become a real competitor to concatenative systems, and would have the built-in advantage of providing implicit control over expressivity.

In the concatenative framework, improvements in selection, voice conversion and interpolation may also lead to more expressive systems. The main challenge on this level is the instability of the synthesis quality in the face of missing data, and the degradation of signal quality incurred by signal modification.

Alternatively, it may be that statistical and concatenative approaches are becoming more similar. Already now, a "hybrid" system that uses HMM-based

statistical models to predict the target for unit selection was among the best ones in the Blizzard Challenge 2007 [43].

A promising avenue for explicit models seems to be opening up on the level of articulatory speech synthesis. Birkholz [44] has proposed a sophisticated system capable of generating intelligible speech from a "score" representation of articulator movements. Studies such as the one by Wollermann and Lasarcyk [45], investigating the use of articulatory speech synthesis for emotional expression, are starting to appear. The tremendous potential of this method is given by the fact that an explicit model of physiological effects of emotions on the articulators could be specified, and its acoustic correlates can be computed through the articulatory model. Things like the tension of the vocal folds, the absorption ratio of the walls of the vocal tract, or the precision of articulation, whose relations to emotions have been described in detail in the psychological literature on emotional speech [46], can be modelled in an explicit way. The main shortcoming of the approach for the non-expert is that there is not yet a fully working Text-to-Speech mode for this system – a considerable amount of expert knowledge is still required to design the articulator score from the text.

One major area where progress is direly needed is the (eplicit or implicit) model of suitable expressions. Manually tuned rules tend to be overly simplified and exaggerated, so that some doubts are due whether purely man-made rules will be most successful. Machine learning methods, on the other hand, depend crucially on suitable models and on large amounts of data. Models must describe not only the acoustics (the predictee), but also the causing factors serving as predictors – emotion, mood, communicative situation, social stance, physiological state, etc. must be expected to interact in producing the acoustic effect, and are likely to be used by a human listener when interpreting the meaning of an acoustic configuration in speech. Large scale data collections such as the one by Campbell [37] will help shed light on these effects if they are supplemented with suitable annotations of predictors.

Another interesting contribution to this area may come from cognitive systems that use the speech synthesizer as an actuator for interacting with their environment, such as Moore's PRESENCE model [47]. Through a feedback mechanism, such a cognitive system can learn how to express itself appropriately in a given situation. Feedback can include, for example, proprioception (e.g., hearing oneself), explicit evaluation and correction by a human user, or success at performing a given task. Situationally adequate "communication strategies" involving expressive speech may emerge from such systems that cannot be predicted today.

5 Summary and Conclusion

This chapter has attempted to trace the development of expressive speech synthesis since its beginnings, with formant synthesis, via early concatenative systems relying on explicit models of expressive prosody and "playback" models, to current challenges in the area, which all seem to address the same issue from dif-

ferent angles: how to combine a high degree of naturalness with a suitable level of control over the expression. Based on the current state of affairs, the chapter also ventures some speculation about future developments, pointing out promising aspects in statistical, concatenative and articulatory synthesis, and the challenge of learning appropriate rules, which may see some interesting "fresh air" from the area of cognitive systems entering the field of speech synthesis.

Acknowledgements

The preparation of this paper has received funding from the EU project HUMAINE (IST-507422), the DFG project PAVOQUE, and from the European Community's Seventh Framework programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE).

References

1. Schröder, M.: Approaches to emotional expressivity in synthetic speech. In Izdebski, K., ed.: The Emotion in the Human Voice. Volume 3. Plural, San Diego, CA (2008)
2. Cahn, J.E.: The generation of affect in synthesized speech. Journal of the American Voice I/O Society **8** (1990) 1–19
3. Murray, I.R., Arnott, J.L.: Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication **16** (1995) 369–390
4. Burkhardt, F., Sendlmeier, W.F.: Verification of acoustical correlates of emotional speech using formant synthesis. In: Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland (2000) 151–156
5. Charpentier, F., Moulines, E.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In: Proc. Eurospeech, Paris, France (1989) 13–19
6. Vroomen, J., Collier, R., Mozziconacci, S.J.L.: Duration and intonation in emotional speech. In: Proceedings of Eurospeech 1993. Volume 1., Berlin, Germany (1993) 577–580
7. Edgington, M.: Investigating the limitations of concatenative synthesis. In: Proceedings of Eurospeech 1997, Rhodes/Athens, Greece (1997)
8. Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., Pardo, J.M.: Analysis and modelling of emotional speech in Spanish. In: Proceedings of the 14th International Conference of Phonetic Sciences, San Francisco, USA (1999) 957–960
9. Heuft, B., Portele, T., Rauth, M.: Emotions in time domain synthesis. In: Proceedings of the 4th International Conference of Spoken Language Processing, Philadelphia, USA (1996)
10. Rank, E., Pirker, H.: Generating emotional speech with a concatenative synthesizer. In: Proceedings of the 5th International Conference of Spoken Language Processing. Volume 3., Sydney, Australia (1998) 671–674
11. Audibert, N., Vincent, D., Aubergé, V., Rosec, O.: Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions. In: Proc. Speech Prosody, Dresden, Germany (2006)

12. Schröder, M.: Can emotions be synthesized without controlling voice quality? Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland (1999) 37–55
13. Schröder, M., Grice, M.: Expressing vocal effort in concatenative synthesis. In: Proceedings of the 15th International Conference of Phonetic Sciences, Barcelona, Spain (2003)
14. Bulut, M., Narayanan, S.S., Syrdal, A.K.: Expressive speech synthesis using a concatenative synthesiser. In: Proceedings of the 7th International Conference on Spoken Language Processing, Denver, Colorado, USA (2002)
15. Iriondo, I., Guaus, R., Rogríguez, A., Lázaro, P., Montoya, N., Blanco, J.M., Bernadas, D., Oliver, J.M., Tena, D., Longhi, L.: Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In: Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland (2000) 161–166
16. Schröder, M.: Emotional speech synthesis: A review. In: Proceedings of Eurospeech 2001. Volume 1., Aalborg, Denmark (2001) 561–564
17. Mozziconacci, S.J.L.: Speech Variability and Emotion: Production and Perception. PhD thesis, Technical University Eindhoven (1998)
18. Schröder, M.: Expressing degree of activation in synthetic speech. IEEE Transactions on Audio, Speech and Language Processing **14** (2006) 1128–1136
19. Mozziconacci, S.J.L., Hermes, D.J.: Role of intonation patterns in conveying emotion in speech. In: Proceedings of the 14th International Conference of Phonetic Sciences, San Francisco, USA (1999) 2001–2004
20. Iida, A., Campbell, N.: Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. International Journal of Speech Technology **6** (2003) 379–392
21. Johnson, W.L., Narayanan, S.S., Whitney, R., Das, R., Bulut, M., LaBore, C.: Limited domain synthesis of expressive military speech for animated characters. In: Proceedings of the 7th International Conference on Spoken Language Processing, Denver, Colorado, USA (2002)
22. Pitrelli, J.F., Bakis, R., Eide, E.M., Fernandez, R., Hamza, W., Picheny, M.A.: The IBM expressive text-to-speech synthesis system for American English. IEEE Transactions on Audio, Speech and Language Processing **14** (2006) 1099–1108
23. Fernandez, R., Ramabhadran, B.: Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In: Proc. 6th ISCA Workshop on Speech Synthesis, Bonn, Germany (2007) 34–39
24. Campbell, N., Marumoto, T.: Automatic labelling of voice-quality in speech databases for synthesis. In: Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China (2000)
25. Wang, L., Chu, M., Peng, Y., Zhao, Y., Soong, F.: Perceptual annotation of expressive speech. In: Proc. 6th ISCA Workshop on Speech Synthesis, Bonn, Germany (2007) 46–51
26. Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S.: Towards emotional speech synthesis: A rule based approach. In: Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, USA (2004) 219–220
27. d'Alessandro, C., Doval, B.: Voice quality modification for emotional speech synthesis. In: Proc. Eurospeech 2003, Geneva, Switzerland (2003) 1653–1656
28. Vincent, D., Rosec, O., Chonavel, T.: Estimation of LF glottal source parameters based on an ARX model. In: Proc. Interspeech, Lisbon, Portugal (2005) 333–336

29. Matsui, H., Kawahara, H.: Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system. In: Proc. Eurospeech 2003, Geneva, Switzerland (2003) 2113–2116
30. Ye, H., Young, S.: High quality voice morphing. In: Proc. ICASSP 2004, Montreal, Canada (2004)
31. Turk, O., Schröder, M., Bozkurt, B., Arslan, L.: Voice quality interpolation for emotional text-to-speech synthesis. In: Proc. Interspeech 2005, Lisbon, Portugal (2005) 797–800
32. Schröder, M.: Interpolating expressions in unit selection. In: Proc. 2nd International Conference on Affective Computing and Intelligent Interaction (ACII'2007), Lisbon, Portugal (2007)
33. Zen, H., Toda, T.: An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In: Proc. Interspeech, Lisbon, Portugal (2005) 93–96
34. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Modeling of various speaking styles and emotions for HMM-based speech synthesis. In: Proc. Eurospeech, Geneva, Switzerland (2003) 2461–2464
35. Miyanaga, K., Masuko, T., Kobayashi, T.: A style control technique for HMM-based speech synthesis. In: Proceedings of the 8th International Conference of Spoken Language Processing, Jeju, Korea (2004)
36. Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., Nakano, Y.: Model adaptation approach to speech synthesis with diverse voices and styles. In: Proc. ICASSP. Volume IV., Hawaii (2007) 1233–1236
37. Campbell, N.: Approaches to conversational speech rhythm: speech activity in two-person telephone dialogues. In: Proc. International Congress of Phonetic Sciences, Saarbrücken, Germany (2007) 343–348
38. Campbell, N.: Developments in corpus-based speech synthesis: Approaching natural conversational speech. Institute of Electronics, Information and Communication Engineers transactions on information and systems **88** (2005) 376–383
39. Trouvain, J., Schröder, M.: How (not) to add laughter to synthetic speech. In: Proc. Workshop on Affective Dialogue Systems, Kloster Irsee, Germany (2004) 229–232
40. Schröder, M., Heylen, D., Poggi, I.: Perception of non-verbal emotional listener feedback. In: Proc. Speech Prosody 2006, Dresden, Germany (2006)
41. Schröder, M.: Experimental study of affect bursts. Speech Communication Special Issue Speech and Emotion **40** (2003) 99–116
42. Ekman, P.: Biological and cultural contributions to body and facial movement. In Blacking, J., ed.: The anthropology of the body. Academic Press, London (1977) 39–84
43. Ling, Z.H., Qin, L., Lu, H., Gao, Y., Dai, L.R., Wang, R.H., Jiang, Y., Zhao, Z.W., Yang, J.H., Chen, J., Hu, G.P.: The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007. In: Proc. Blizzard Challenge, Bonn, Germany (2007)
44. Birkholz, P.: Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In: Proc. Interspeech, Antwerp, Belgium (2007)
45. Wollermann, C., Lasarcyk, E.: Modeling and perceiving of (un-)certainty in articulatory speech synthesis. In: Proc. 6th ISCA Speech Synthesis Workshop, Bonn, Germany (2007) 40–45
46. Scherer, K.R.: Vocal affect expression: A review and a model for future research. Psychological Bulletin **99** (1986) 143–165
47. Moore, R.K.: Spoken language processing: Piecing together the puzzle. Speech Communication **49** (2007) 418–435