# Rule-Based Voice Quality Variation with Formant Synthesis

*Felix Burkhardt*

Deutsche Telekom Laboratories
Ernst Reuter Platz 7, 10587 Berlin
Germany
felix.burkhardt@telekom.de

## Abstract

We describe an approach to simulate different phonation types, following John Laver's terminology, by means of a hybrid (rule-based and unit concatenating) formant synthesizer. Different voice qualities were generated by following hints from the literature and applying the revised KLGLOTT88 model. Within a listener perception experiment, we show that the phonation types get distinguished by the listeners and lead to emotional impression as predicted by literature. The synthesis system and its source code, as well as audio samples can be downloaded at *http://emoSyn.syntheticspeech.de/*.

## 1. Introduction

The naturalness of synthetic speech is still an open issue. Although, thanks to non-uniform unit selection synthesis [1], the generation of high quality speech within a limited domain has made great progress in the last decade, the simulation of different speaking styles remains an open question.

The ultimate goal of speech synthesis, to create a system unlimited by text, language, speaker characteristics or speaking style, can clearly not be reached by a system that is unaware of the underlying speech mechanisms.

Formant synthesis, especially if complex voice source models are included, incorporates direct methods to simulate different phonation types. Parametric techniques, due to insufficient models, which to a great part concerns source filter interaction, still sound very unnatural and didn't perceive much attention in the past decade. Nonetheless, we believe that the need for more flexible models, better understanding of speech mechanisms and the advances in hardware and algorithms will lead to a renaissance.

One aspect of formant synthesis renaissance is its use as a signal generation model in pseudo-articulatory synthesis, i.e. the speech production is modeled on a higher level, e.g. by articulatory movements, which get translated via 3D model into formant tracks.

The formant synthesis approach can be also used in combination with data-driven synthesis, i.e. when the units are coded as formant parameters as shown in [2].

We introduced a formant synthesis framework based on a hybrid concatenating and rule-based approach similar to [2] in [3]. It is based on the freely available Sensyn Klatt-style synthesizer (http://www.sens.com) and can be downloaded at http://emoSyn.syntheticspeech.de/.

In this paper, we describe the rule-based simulation of phonation types in detail, by presenting the formulas which are used to change the default values in the glottal source model parameters.

After discussing related work in the next section, we describe the glottal source model parameters. We then give a detailed description on the manipulations done for each of the phonation types. A perception experiment was conducted in order to test the perceptive relevance of the changes and its ability to induce affective impression.

## 2. Related Work

Laver [4] did a characterization of different phonation types and described modal, breathy, whispery, creaky, tense and lax voice, which play an important role in affect display, and is still seen as a widely accepted reference system, see e.g. [5], p. 193.

The topic of voice source modeling is of course as old as speech synthesis. Parametric synthesis methods like LPC- or formant synthesis are based on a source filter model of speech production [6] and modeling the source is an important part with respect to natural sounding speech. Although for intelligible speech, a simple pulse train is sufficient to model the voice source, more complex models are needed. Firstly, source filter interaction needs to be regarded for, i.e. the source is not completely independent of the filter characteristics and should change with the phonemes. Secondly, in order to model speech characteristics like male, female or children's voices, phonation types, speaking style and emotional expression, appropriate excitation signals are needed.

Klatt et al introduce their KLSYN88 synthesizer with a specific glottis model to increase naturalness [7]. They include support for the Liljencrants Fant model [8] parameters, a widely used five parameter model to describe voice source periodic signals and phonation types, as done e.g. by Chasaide and Gobl, described in [9].

Carlson et al extend their existing formant synthesizer OVE by a glottis model using the LF parameters and conduct voice quality re-synthesis experiments [10]. Karlsson also uses the LF model to synthesize sonorant, strained and breathy voice [11]. Epstein [12] used the LF-Model to model pathological voices and are successful in re-synthesis experiments. Hermes describes a method to synthesize breathy vowels by combining low pass filtered pulses with high pass filtered noise bursts [13].

Gobl and Chasaide did many studies on voice quality and its connection to affective speech, a comprehensive overview is given in [5]. Beneath basic emotions they also investigate moods and speaker attitudes. A references stimulus spoken in a modal voice was recorded and inverse filtered in order to separate source and filter characteristics. Using the KLSYN88 synthesizer, the authors simulated several phonation types by altering the glottis parameters and related the outcome to emotional impression by means of a perception experiment.

## 3. Parameters of the Klatt formant synthesizer

In Figure 1, the parametrization of the glottal flow and it's first derivative is shown. It consists of $Ee$, the maximal negative amplitude, it's time point $te$, $tp$, the time of the first zero crossing, and $Ta$, a parameter that correlates with spectral dampening. Klatt et al introduced in [7] a revised source model for their synthesizer which includes, derived partly from the Liljencrants Fant parameters, the following parameters.

- $F_0$: the fundamental frequency.
- AV: amplitude of voicing, relates to $Ee$.
- OQ: the open quotient: the relation between open phase and period duration.
- AH: amplitude of aspiration, not described by the LF model.
- TL: spectral tilt: a dampening of the higher frequencies, relates to $Ta$.
- FL: $F_0$-flutter: a jitter simulation that shifts period duration randomly, affects more than one period.
- DI: diplophonic double pulsing: shift and amplitude change of every first period in a pair, used to simulate laryngealized voice, affects more than one period.

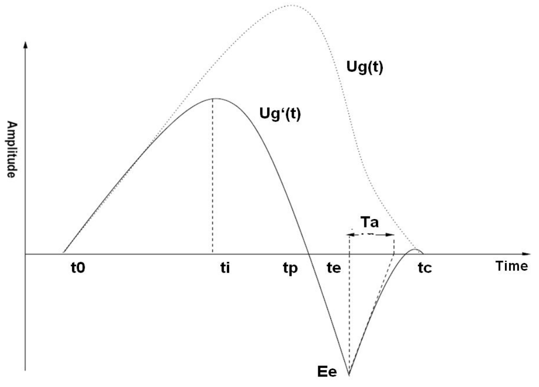

Figure 1: Parametrization of the glottal flow by the LF model

## 4. Modeling phonation types

The acoustic correlates of different phonation types were described among others by [4, 7, 9, 11, 13, 5]. These hints were used to modify the source signal parameters, the amplitude and intensity values as well as formant bandwidths.

### 4.1. Breathy voice

Breathiness can be generated with a strength between 0 and 100. The parameters for open quotient (OQ), spectral damping (TL) bandwidth of first formant (B1), amplitude of voiced part (AV) and amplitude of noisy part (AH) get modified following these formulas:

$$\begin{aligned}
OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/100\\
TL_{add} &= (TL_{max} - TL_{glob}) * rate/100\\
B1_{add} &= (B1_{max} - B1_{glob}) * rate/100\\
AV_{sub} &= 6 * rate/100\\
AH_{add} &= (AMP - 3) * rate/100
\end{aligned}$$

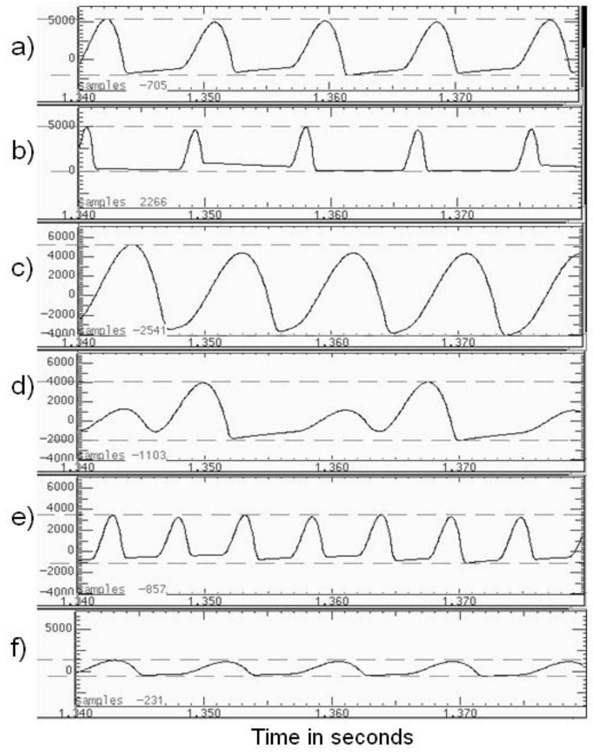

Figure 2: Source signals for different phonation types. a: modal, b: tense, c: breathy, d: creaky, e: falsetto and f: whispery voice

$OQ_{add}$, $TL_{add}$, $B1_{add}$ and $AH_{add}$ are the summands to the respective parameters and $AV_{sub}$ the subtrahend to $AV$. $OQ_{glob}$ and $TL_{glob}$ are the global constants to $OQ$ and $TL$, $AMP$ is the amplitude value of the frame and $rate$ the degree of breathiness in percent. $OQ_{max}$ (100 %), $TL_{max}$ (24 dB) and $B1_{max}$ (250 Hz) denote the maximal values after the modification.

Furthermore we simulate the poles that arise from tracheal coupling (following [14]). The frequencies of the nasal pole and zero (FNP and FNZ) get set to 550 Hz and the tracheal pole and zero (FTP and FTZ) get set to 2100 Hz (see [14], page 68. The bandwidths of the poles get narrowed using the following formulas:

$$\begin{aligned}
BTP_{sub} &= (BTP_{glob} - BTP_{min}) * rate/100\\
BNP_{sub} &= (BNP_{glob} - BNP_{min}) * rate/100
\end{aligned}$$

$BTP_{sub}$ and $BNP_{sub}$ are the respective subtrahends, $BTP_{glob}$ and $BNP_{glob}$ (180 and 100 Hz) the constant default values, and $BTP_{min}$ and $BNP_{min}$ (30 Hz) the respective minimal values. The amplitude track of the of the voiced component for breathy excitation is shown in figure 2 c). One can clearly see the almost sinusoidal excitation form the results from the damping of higher frequencies.

### 4.2. Tense voice

*Tense voice* is not really a phonation type, because also supralaryngeal components are effected by it. [4] describes a phonation type with similar features named "*harsh voice* ", and uses *tense voice* as a global setting. We use it here, as other authors, e.g. [15]), because it correlates strongly with the emotion dimension of arousal.

The parameters AV, OQ, TL and B1-5 get modified as follows:

$$\begin{aligned} AV_{add} &= 6 * rate/100 \\ OQ_{sub} &= (OQ_{glob} - OQ_{min}) * rate/100 \\ TL_{sub} &= TL_{glob} * rate/100 \\ B1_{sub} &= (B1_{glob} - B1_{min}) * rate/100 \\ &\dots \\ B5_{sub} &= (B5_{glob} - B5_{min}) * rate/100 \end{aligned}$$

The amplitude of the voiced part gets raised by up to 6 dB. Th open quotient gets decreased to the minimal meaningful value ($OQ_{min} = 10$) and the spectral damping gets turned off in the extreme case of $rate = 100$.

In order to account for the weaker damping caused by the tenseness of the tissue, all formant bandwidths get reduced up to a minimal value ($BX_{min}$ =30-60 Hz).

A picture of the periodic component for tense voice with a rate of 70 % is displayed in figure 2 b). One can see clearly the, compared to modal voice, shorter open phase and steeper flanks of the signal.

### 4.3. Whispery voice

Whispery voice gets simulated mainly by a replacement of the voice part AV by glottal noise AH:

$$\begin{aligned} AV_{sub} &= AV * rate/100 \\ AH_{add} &= AV * rate/100 \end{aligned}$$

Furthermore the open quotient gets strengthened, the spectral dampening raised and the bandwidths of all formants get broadened:

$$\begin{aligned} OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/100 \\ TL_{add} &= (TL_{max} - TL_{glob}) * rate/100 \\ B1_{add} &= (B1_{max} - B1_{glob}) * rate/100 \\ &\dots \\ B5_{add} &= (B5_{max} - B5_{glob}) * rate/100 \end{aligned}$$

The remaining periodic parts of the excitation signal (without noise) are shown in 2 f), with a degree of 30 %. The amplitude is clearly reduced and the track more sinusoidal.

### 4.4. Creaky voice

Creaky voice is characterized mainly by a more irregular and very low fundamental frequency. There might by diplophonic double pulsing, which was already in the second version of the Klatt synthesizer accounted for ([7]) (parameter DI).

With diplophonic double pulsing, every first period gets moved nearer to the second and damped with respect to its amplitude. Like this, both periods get perceived as one and the fundamental frequency halved in effect.

As a rule, creaky voice goes with a short opening phase and steep flanked pulses ([4], p. 125 and [14], p. 7), stemming from the relatively strong medial compression and adductive tenseness in the larynx.

On the other hand, creaky voice is quite variable ([9], p. 450), and often goes with emotions that are characterized by a low degree of arousal, like e.g. sadness ore boredom. If the subglottal pressure is rather low and the muscular tenseness not so strong, the flanks can also be less steep and higher frequencies contain less energy.

In order not to disagree with Laver's terminology, we call this kind of phonation, as did [5], in the following '*lax creaky voice*'. It was noticed in our emotional database ([16]) and is described further in [17], p. 121.

$$\begin{aligned} DI &= rate \\ OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/200 \\ AV_{sub} &= 6 * rate/100 \\ B1_{add} &= (B1_{max} - B1_{glob}) * rate/100 \end{aligned}$$

We model lax creaky voice with an elongated opening phase, up to half the maximum value, and a broader bandwidth of the first formant.

It goes along with relatively weak energy in the area of the fundamental component ([14]), which we take into account by lowering the amplitude of the voiced excitation AH.

The signal of lax creaky voice with a degree of 70 % is displayed in figure 2 d). Each first period is dampened and a bit displaced. The open phase is elongated and the flanks less steep.

### 4.5. Falsetto voice

Falsetto is acoustically characterized by a high fundamental frequency and a strong dampening of higher frequencies. We take this into account by a maximum rise of $F_0$ by 100 % and a rise of spectral dampening by up to 24 dB.

Because partly the total closure of the glottis is prevented ([4], p. 118), we elongate the opening phase.

In order to account for irregularities resulting from the incomplete glottal closure, the flutter parameter FL is set to the rate of falsetto.

$$\begin{aligned} F0_{add} &= F0 * rate/100 \\ OQ_{add} &= (OQ_{max} - OQ_{glob}) * rate/100 \\ TL_{add} &= (TL_{max} - TL_{glob}) * rate/100 \\ FL &= rate \end{aligned}$$

We didn't model the added noise parts coming from the missing closure, because they are rather weak with low subglottal pressure ([4], p. 119). The periodic excitation with 70 % falsetto is depicted in figure 2 e). The high fundamental frequency is clearly visible.

## 5. Perceptual evaluation

This work was done within the scope of a broader study concerned with the phonetic correlates of emotional speech. In order to test the perceptual relevance of the simulated phonation types, we carried out a listening experiment.

We did not ask the listeners to recognize Laver's phonation types though, but to identify emotional vocal expression, because firstly the overall subject of the study were the acoustical correlates of affective speech and secondly naive listeners might be unfamiliar with Laver's terms. This was done also by other authors [5]. Nonetheless, from the fact that the phonation types were distinguished and there are many studies that deal with the correlation of voice quality and emotion expression, we take this as an indication that the simulation was to a degree successful.

A set of target sentences were copy-synthesized and then varied with respect to their voice quality by using the formulas described in section 4 with the following rates: modal, falsetto, tense and breathy: 70%, and creaky: 30%. As whispery voice seemed less important for emotional expression and the stimulus set was already quit large, we didn't include it in the experiment.

We carried out a forced choice listening experiment with 30 participants, gender weighted, all of them German native speakers. Each stimulus had to be assigned to either neutral, anger, joy, sadness, frightened or bored speech. In figure 3, we display the results for the different phonation types that reached statistical significance.

For breathy as well as for creaky voice, sadness and boredom were mostly favored by the listeners, which goes along with results presented in the literature.

With falsetto voice, the listeners voted on the one hand for a frightened expression, which goes along with the literature, on the other hand sadness reached quite a high voting. Probably the listeners detected a kind of sadness with high arousal, a kind of "whiny" sadness.

Modal voice did not reach a statistical significant statement, which of course is desired in the scope of this study, with the exception that it clearly does not sound angry.

As predicted by the literature, tense voice promotes an angry speaker impression. The fact that sadness was also rated high for tense phonation, can perhaps be explained by the fact that all phonation types except modal lead to a sad impression. We conclude that the category "sadness" is a too raw concept and should be divided into "quite sadness" (with low arousal) vs. "whiny sadness" (with high arousal).

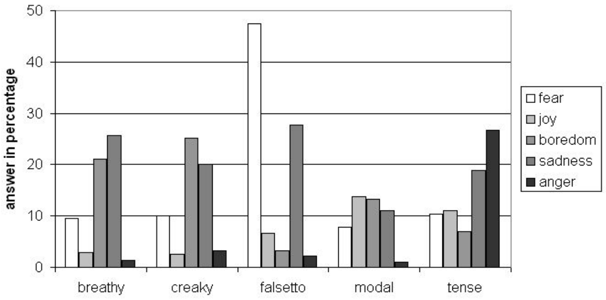

Figure 3: Results of the evaluation for different phonation types

## 6. Conclusions

In this study, we experimented on the simulation of different phonation types by means of a formant synthesizer. We showed in detail how we generated the phonation types with the LF model following hints from Laver's work and related literature. Within a listener perception experiment, we could show that the phonation types were indeed distinguished by the listeners and lead to emotional impression as predicted by literature. Although formant synthesis played a minor role in the research on speech synthesis of the last decade, we believe that the need for more flexible models, better understanding of speech mechanisms and the advances in hardware and algorithms will lead to a renaissance, both as part of pseudo-articulatory synthesis or as part of a formant-driven non-uniform unit concatenation system. The synthesis system and its source code, as well as audio samples can be downloaded at *http://emoSyn.syntheticspeech.de/*.

## 7. Acknowledgments

This work was partially funded by the EU FP6-IST project HUMAINE (Human-Machine Interaction Network on Emotion) and by a project funded by the DFG (German Research Community), No. Se 462/3-1. The authors further like to thank Christer Gobl and Ailbhe Ni Chasaide, which did great work on the fundamentals of voice source synthesis, for their support.

## 8. References

[1] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "Atr – talk speech synthesis system," *Proc. ICSLP92*, vol. 1, pp. 483–486, 1992.

[2] R. Carlson, T. Sigvardson, and A. Sjölander, "Data-driven formant synthesis," KTH, Stockholm, Sweden, Progress Report 44, 2002, http://www.speech.kth.se/qpsr/tmh.

[3] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proc. of the ISCA Workshop on Speech and Emotion*, 2000, pp. 151–156.

[4] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

[5] C. Gobl and A. Ni Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[6] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 's-Gravenhage, 1960.

[7] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice qualtity variations among female and male talkers," *JASA*, vol. 87, no. 2, pp. 820–856, 1990.

[8] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmiss. Lab. - Q.Prog.Stat.Rep*, vol. 4, pp. 1–13, 1985.

[9] A. Ni Chasaide and C. Gobl, "Voice source variation," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds. Blackwell, Oxford, 1997.

[10] R. Carlson, G. Granström, and I. Karlsson, "Experiments with voice modeling in speech synthesis," *Speech Communication*, vol. 10, pp. 481–489, 1990.

[11] I. Karlsson, "Modelling voice variations in female speech synthesis," *Speech Communication*, vol. 11, pp. 491–495, 1992.

[12] M. Epstein, B. Gabelman, N. Antonanzas-Barroso, B. Gerratt, and J. Kreiman, "Source model adequacy for pathological voice synthesis," *Proc. iICPhS 99*, 1999.

[13] D. J. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Communication*, pp. 497–502, 1991.

[14] D. H. Klatt, *KLATTALK, The Conversion of English Text to Speech*. unpublished, 1990, ch. 3: Description of the Cascade/Parallel Formant Synthesizer.

[15] C. Gobl and A. Ni Chasaide, "Acoustic caharcteristics of voice quality," *Speech Communication*, vol. 11, pp. 481–490, 1992.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech 2005, Lisbon*, 2005.

[17] F. Burkhardt, "Simulation emotionaler sprechweise mit sprachsyntheseverfahren," Ph.D. dissertation, TU-Berlin, 2001.