

A Model-Based Experiment towards an Emotional Synthesis

Jonas Lindh

Department of Linguistics
Göteborg University

Abstract

The most successful methods to induce emotions on state of the art unit selection speech synthesis have been built by switching speech database depending on the desired emotion. These methods require a substantial increase of memory compared to a single database and are computationally slow. The model-based approach is an attempt to reshape a neutrally recorded utterance (comparable to the desired output from a modern unit selection system) into simulating a recorded model of a desired emotion.

Factors for manipulation of duration, amplitude and formant shift ratio are calculated by comparing the recorded neutral utterance with three recorded, basic emotional models in accordance with discrete emotion theory – sadness, happiness and anger. F0 (regarded as the intonation) is copied from the model and is then imposed on the neutrally recorded utterance.

The evaluation of the experiment shows that subjects easily categorize discrete emotions in a forced choice. They also grade the resynthesized emotional quality from the neutrally recorded utterance almost equally high as the naturally recorded models for the male voice. The female voice created more difficulties and contained more synthetic artifacts, i.e. it was judged to have a lower quality than the recorded models.

Background and Introduction

Creating emotional synthesis has been a research area for quite some time. Formant speech synthesis is easily distinguished from human speech not only because of the underdeveloped naturalness, but also due to the lack of expressiveness. Several attempts to implement emotions in formant synthesis have taken place (Cahn, 1988; 1989; 1990; Carlson et al., 1992).

When dealing with emotional content in speech the point of departure is almost always the neutral utterance. What is neutral speech, i.e. speech without emotions? Normally, neutral speech is thought of as a carrier being modulated to reveal the emotions being communicated. Such a concept is rather useful when it

comes to synthesizing expressive speech. One simply treats the relationship in a hierarchy where the abstract underlying expression is neutral and the surface expressions are the emotions we want to induce, in this case the basic emotions from discrete emotional theory - anger, sadness and happiness (Levenson, 1994; Laukka, 2004; Tatham & Morton, 2004; Narayanan & Alwan, 2004).

A modern state of the art unit selection speech synthesis normally produces a sentence as neutrally as possible in order to avoid undesired side effects or miscommunication. Neutral in this case means near monotone or containing as few speech fluctuations as possible. This is not always desirable when it comes to for example dialogue systems. To be able to compare whether a system succeeds in expressing a certain emotion or desire, it is obviously also important to study how well people in general succeed in communicating emotions.

The development of conversational systems has increased, meaning that understandable, neutral synthetic speech is barely acceptable anymore. Some success has been reached, but the best ones still depend too much on stored data, including a separate emotional speech database. (Bulut et al., 2002)

The most successful attempts to synthesize emotions have been built by using additional speech databases containing only recordings representing specific emotions uttered (this applies to concatenative/unit selection synthesis systems). The system has to be able to switch database when a specific emotion is desirable. The system must perhaps also use different algorithms/analyses for the different databases since the acoustic content might differ significantly. The databases needed for such a system also mean a substantial increase of data to choose from. A simpler and computationally more efficient method is to induce rules for expressive speech and resynthesize an utterance produced by the system.

Nowadays, most unit selection systems are created by recording a single professional speaker and then using specified parts (nor-

mally diphones as basic element) of the utterances to concatenate new ones. This normally means that a professional speaker must be available to be recorded for emotional utterances of different lengths. If these recordings are used as models, they will then hopefully not differ more from the utterances that will be produced by the system than there will be differences for a specific speaker.

The desire for creating a simpler way of inducing emotions in unit selection synthesis based on rules have been proposed by for example Murray et al. (2000) and Zovato et al. (2004). However, in this paper an experiment using models to calculate differences between a neutral and an emotional utterance is presented and tested. The results show both difficulties and promising results, which are then discussed concerning how to find ways to induce emotions in synthesis. If emotions are to be created by a system they cannot be expected to outperform the communication of emotional content from recorded models.

Method

Two speakers, one female and one male, were recorded uttering the same sentence ("Jag har inte varit där idag") in four different expressive styles: natural, sad, happy and angry. The recordings were made in a studio environment using a high quality microphone. The speakers were told to first consider how to express the emotions in speech concerning duration, amplitude and intonation. They were then told to express the emotions as clearly as possible while recorded, even though the semantic content did not suggest a specific emotion.

Each recorded emotion was then used, both as a model to induce the specific emotion in the neutrally recorded utterance as well as a reference against which the resynthesized speech should be compared. If one uses the same speaker and the calculated differences from the same utterance with different emotions one should be able to resynthesize at least the specific parameters correctly. Six subjects finally evaluated the results by categorizing and grading the neutral recording, the recorded models and three resynthesized objects for the two speakers, i.e. fourteen utterances of the same sentence.

The model-based approach

The approach described and tested in this paper is similar to the rule-based idea that is described in Zovato et al. (2004) and Murray et al. (2000), except that the rules are based on interactive calculations compared to models. The model calculations are also applied to the complete utterances and not applied to specific units (i.e. syllables or diphones etc.).

A state of the art unit selection synthesis attempts to sound as natural and neutral as possible. If the voice used in the system is recorded to produce models of emotions, the neutrally produced output can be seen as the underlying neutral representation. The representation can then be compared to the produced models to be able to calculate variations for specified parameters.

The aim of the model-based approach is to approach the recognition rate for the models themselves and keep naturalness. The limitations are obvious, when stretching and changing too much, PSOLA will create synthetic artifacts.

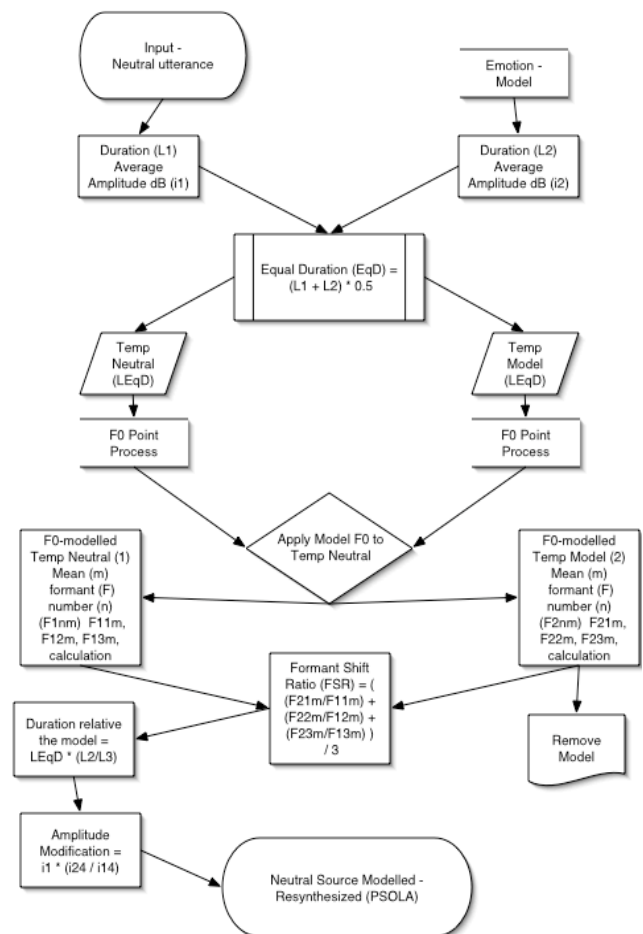


Figure 1. Flow chart showing the script procedure for the model based experiment.

Figure 1 shows how a neutral utterance and a model is compared and the neutral utterance finally resynthesized. First, the neutral utterance and model duration and average amplitude are calculated. Equal duration is then calculated for the objects. Pitch tier objects are then created after point processing the framed fundamental frequency values. A point-processed object is a sequence of points (t_i) in time, defined on a domain $[t_{min}, t_{max}]$. The index (i) runs from 1 to the number of points. The points are sorted by time (i.e. $t_{i+1} > t_i$). Points are generated along the entire time domain of the pitch tier, because there is no voiced/unvoiced information, then the F0 contour is linearly interpolated. This means that one can easily exchange the point processed signal tier from one object to another, thus cloning the intonation (Boersma & Weenink, 2005). The formant shift ratio is then calculated for the first three formants and manipulated. Finally the duration (relative to the model) and the average amplitude is modified and resynthesized.

Results and Discussion

The result of the modulations was calculated by comparing averages and standard deviations for the resynthesized objects and the models.

Table 1. Model and modified parameter values for the male voice

Male voice	F0 mean	F0 std	Ampl (dB)	F1 mean	F2 mean	F3 mean
Neutral	95	24	68	519	1482	2644
Sad Model	153	18	69	405	1300	2592
Resynth	148	16	69	512	1508	2517
Happy Model	133	52	75	528	1464	2602
Resynth	133	46	72	522	1451	2629
Angry Model	84	8	70	517	1367	2672
Resynth	83	5	68	507	1452	2629

Table 2. Model and modified parameter values for the female voice

Female Voice	F0 mean	F0 std	Ampl (dB)	F1 mean	F2 mean	F3 mean
Neutral	172	17	70	573	1670	2687
Sad Model	328	73	67	587	1535	2783
Resynth	311	25	68	610	1651	2681
Happy Model	358	119	77	707	1661	2709
Resynth	349	107	73	608	1734	2767
Angry Model	250	53	77	638	1658	2649
Resynth	236	52	74	614	1689	2686

As can be observed in Table 1 the F0 values are modified fairly well compared to the models. The formant shift ratio should be individualized to each formant and not changed depending on the general averages from the first three. For the female voice (table 2) the neutral recording contained some traces of creakiness, which led to some failure in the F0 analysis and thereby also the resynthesis. Generally, the values approach the model's.

Evaluation Test

Seven subjects with normal hearing and some previous experience of listening to synthesized speech (six employees and one student at the department of linguistics) performed an evaluation. In the evaluation the subjects listened to sixteen samples, eight male and eight female. The samples were the neutral utterance plus the three recorded models of the same sentence and the three resynthesized samples. When hearing the samples the subjects had to categorize each sample belonging to one of the four categories neutral, happy, sad and angry. After categorizing they had to grade the confidence level of their categorization from 1 to 5 (absolutely confident). They finally had to grade the naturalness, meaning a score between very synthetic (1) to sounding like a recorded voice (5). The average results calculated from all subjects in Table 3 and 4 below.

Table 3. Average results from the categorization and grading (1-5) by seven subjects of the male voice.

Male voice	Neutral	Sad	Happy	Angry	Natural
Neutral	4.7				4.3
Sad Model		4.3			3.8
Resynth		4.2			2.7
Happy Model			4.8		4.7
Resynth	0.7		3		3.5
Angry Model				3.67	4.33
Resynth				3.5	3.5

The average naturalness score for the four resynthesized male samples is 3.37, while the overall average for the recorded models is 4.29. Whether this decrease in naturalness is an acceptable has not been investigated. Categorizing the male samples created no problems except one uncertain exception (0.7 happy-neutral). This means that there is a trade-off between naturalness and an computationally cheap method.

Table 4. Average results from the categorization and grading (1-5) by seven subjects of the female voice.

Female voice	Neutral	Sad	Happy	Angry	Natural
Neutral	0.7			2.8	4
Sad Model		4.5			3.5
Resynth	0.8	1.8			1.7
Happy Model			4		3.5
Resynth	0.3		1.2	1	2
Angry Model				5	4.7
Resynth	0.5			2.8	2.8

The female voice created more difficulties. The samples contained more synthetic artifacts, which was detected by the listeners. The average naturalness score for the resynthesized samples is 2.62. Since the categorization as well as the grading was worse here, it is likely that the synthetic low quality of the output made categorizing more difficult. This might also be an example of what happens when bad models are created (see table 2).

Conclusions and Further Developments

Categorizing discrete emotions does not seem to be a problem. The difficulty certainly increases as quality degrades. Female voices turned out to be more difficult to resynthesize without degrading naturalness. More research is needed to be able to make a well-grounded comparison. One problem may be individual voice variation.

The purpose of a model-based approach is to be able to induce discrete emotions on a neutrally uttered sentence, produced by a state of the art unit selection system. By comparing well-formed models from one individual speaker, characteristics such as intonation, F0-changes, formant shift ratios and amplitude can be calculated and induced on a neutrally uttered sentence successfully.

More research on which segment level (syllable, diphone etc.) the calculations and inducing should be done is desirable for the future. There also remains several questions regarding what a model should look like and which parameters that really should be modified to reach the model-based goal.

References

Boersma P. & Weenink D. (2005) Praat: doing phonetics by computer (Version 4.3.07)

- [Computer program]. Retrieved March 31, 2005, from <http://www.praat.org/>
- Bulut M., Narayanan S., Syrdal A. (2002) Expressive speech synthesis using a concatenative synthesizer. In Proc. ICSLP (Denver)
- Cahn J. E. (1990) The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society*, Volume 8. 1-19.
- Cahn J. E. (1989) Generation of Affect in Synthesized Speech. *Proceedings of the 1989 Conference of the American Voice Input/Output Society*. Newport Beach, California. Pages 251-256.
- Cahn J. E. (1988) From Sad to Glad: Emotional Computer Voices. *Proceedings of Speech Tech '88, Voice Input/Output Applications Conference and Exhibition*. New York City. Pages 35-37.
- Carlson R., Granström B., Nord L. (1992) Experiments with emotive speech - acted utterances and synthesized replicas. *Proc. ICSLP'92*, pp. 671-674
- Laukka P. (2004) *Vocal Expression of Emotion. Discrete-emotions and Dimensional Accounts*, Acta Universitatis Upsaliensis, *Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 141. 80pp. Uppsala.
- Levenson R. W. (1994) Human emotion: A functional view. In P. Ekman & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp.123-126). New York: Oxford University Press
- Murray I. R., Edgington M. D., Campion D., Lynn J. (2000) Rule-based emotion synthesis using concatenated speech, In *SpeechEmotion-2000*, 173-177.
- Narayanan S., Alwan A. (2004) *Text to Speech Synthesis: New Paradigms and Advances*. Prentice Hall PTR, IMSC Press Multimedia Series.
- Tatham M. & Morton K. (2004) *Expression in speech : analysis and synthesis*. Oxford [England] ; New York : Oxford University Press.
- Zovato E., Pachiotti A., Quazza S., Sandri S. (2004) Towards emotional speech synthesis: a rule based approach. *Workshop Proceedings, 5:th ISCA Speech Synthesis Workshop*, Carnegie Mellon University, Pittsburgh (US).