THE PERCEPTUAL IMPORTANCE OF SELECTED VOICE QUALITY PARAMETERS

Gudrun Klasmeyer
Institute for Communication Science, TU Berlin,
Sekr. EN 8, Einsteinufer 17, 10587 Berlin, Germany
klasmeyer@kgw.tu-berlin.de

ABSTRACT

It is well known, that personal voice qualities differ in the speakers use of temporal structures, F0 contours, articulation precision, vocal effort and type of phonation. Whereas temporal structures and F0 contours can be measured directly in the acoustic signal and conclusions about articulation precision can be made from the formant structure, this paper focuses especially on vocal effort and type of phonation. These voice quality percepts are a combination of several acoustic voice quality parameters: the glottal pulse shape in the time domain or damping of the harmonics in the frequency domain, spectral distribution of turbulent signal components and voicing irregularities. In an investigation on emotionally loaded speech material it could be shown, that the named acoustic parameters are useful for differentiating between the emotions happiness, sadness, anger, fear and boredom [1].
The perceptual importance of the above acoustic parameters is investigated in perception experiments with synthetic and resynthesized speech.

1. DATABASES

For the analysis, 10 short sentences are used which could appear in everyday communication in all emotional situations. The sentences were spoken with 6 different emotions by 3 actors. The recognizability and naturalness of the emotions was evaluated in a listening test with 60 naive listeners. Only those stimuli with very high recognition scores (80%) were used for the analysis.
For the perception experiments (Experiment A and Experiment B, Part 1 & 2) stimuli were partly synthesized and partly resynthesized from the original material after manipulating selected acoustic parameters.

2. SIGNAL ANALYSIS

As stated above, this paper focusses especially on the voice quality percepts vocal effort and type of phonation.
Vocal effort can still be perceived from the sound quality regardless of how loud they actually are presented. Objective measures of vocal effort in the acoustic signal are glottal pulse shapes or spectral damping of the harmonics in the frequency domain [2]. The glottal pulses can (with some limitations) be calculated from the acoustic speech signal. The method and the problems that might appear are described in Klasmeyer & Sendlmeier [3]. Steep pulses are usually a sign of high vocal effort. An additional estimation of vocal effort can be made from an energy measurement in a vowel and a pre- or antecedent voiceless fricative: The possible loudness range of voiceless fricatives is more limited than the possible loudness range of vowels, therefore the fricative energy can be used as (a vague) reference. The energy difference is measured in dB.
From glottal pulse shapes and additional measurements of the spectral distribution of harmonic and noise components the **type of phonation** can be classified within the phonetic categories defined by Laver [4]: breathy

phonation, whispery phonation, modal voicing, creak (glottal fry) and falsetto. The spectral distribution of harmonic and noise components is investigated on a phoneme base. The phoneme based investigation is necessary to separate voice quality information from the linguistic content in the analysis procedure. In creaky phonation and in all segments with steep F0 changes, the FFT is not precise enough to describe the signals constitution of harmonic and noise components. Therefore a vowel segment without formant movements is chosen (phoneme /a/) by the help of a wideband spectrogram. Four different frequency bands are filtered from the segment: the very-low-band, which only contains the fundamental frequency, the low-band (0-1kHz), the mid-band (2.5-3.5kHz) and the high-band (4-5kHz). The energy within the frequency bands is calculated and devided by the energy of the whole segment. In order to decide, whether the band contains mainly harmonics of the fundamental frequency or turbulent noise, the Hilbert envelope of each band is calculated. (The method is described in Michaelis [5].)

3. PERCEPTION EXPERIMENTS

Two experiments were designed: In the first experiment (Experiment A) the perceptual effect of glottal pulse shapes typical for specific emotions is investigated. The stimuli are presented in pairs to judge the perceptual differences.
In the second experiment (Experiment B, Part 1) synthetic stimuli are used. The stimuli represent a wider range of physiologically possible voice characteristics like the spectral distribution of harmonic and noise components or voicing irregularities. These stimuli are also presented in pairs to judge the perceptual differences on a rating scale. In the second part of the experiment, the synthetic voice signals are filtered with the formants from a completely voiced sentence to achieve a speechlike quality of the stimuli.

Experiment A

According to the results derived from the analysis of emotional speech data, six **glottal pulse shapes** were chosen. Fundamental frequency and loudness were normalized amoung the stimuli to investigate the perceptual effect of the pulse shape only. Glottal pulses for the emotions anger, happiness, boredom and neutral could be calculated by inverse filtering of the emotional speech data. The pulses differ in the spectral damping of the harmonics. While anger and happiness show little damping, the pulses for the neutral reference and especially for boredom have less energy in higher frequency regions. Glottal pulses for sadness had to be created synthetically from a sine wave and two harmonics with an amplitude ratio of 16:4:1 from the lowest to the highest frequency. The spectral damping of this synthetic pulse is high in comparison to the inverse filtered pulses. For fear the same glottal pulse was used with additional band noise in the region 1-5 kHz and the noise amplitude similar to that of the fundamental frequency. Before presenting the stimuli in a listening test, all pulses were filtered with the formants of the phoneme /a/.

The duration of each stimulus was 2 sec with a pause of 1 sec between the stimuli and a pause of 3 sec between the pairs. The naive listeners judged the similarity of the stimuli within each pair on a linear rating scale of 6 cm length. Some listeners tended to use only the middle of the rating scale. Therefore the judgements were normalized setting the minimum similarity to 0 and the maximum similarity to 1 for each listener. For each stimulus pair the average similarity and standard deviation was calculated and the pairs were arranged in an order of decreasing similarity.

Results

None of the pairs were judged similar. Those pairs in which the noisy glottal pulse signal (emotion fear) was compared to completely harmonic pulses were judged as being most different. The noisy components mask the harmonic components almost completely, the voice quality can be described as whispery or almost voiceless. The pairs anger/boredom, sadness/neutral, sadness/anger, sadness/happiness and sadness/boredom were

placed in the middle of the rating scale. In general, the spectral damping of the stimuli in these pairs was very different. The pairs neutral/boredom, neutral/ happiness, happiness/anger, anger/neutral and happiness/neutral were judged as being most similar. This is in accordance with their physical similarity.

The glottal pulse shapes taken from the emotional speech database do not cover the complete range of possible glottal pulse shapes, but still the differences in the sound quality could be perceived clearly by naive listeners. This result encouraged a second experiment.

Experiment B - Part 1

In Experiment B synthetic stimuli were used which cover a wider range of physiologically possible voice characteristics. The experiment was designed to evaluate the perceptual effect of the acoustic variables: **spectral damping of harmonics, amplitude of turbulent components, voicing irregularities and falsetto (high fundamental frequency)**.

A "male voice" was represented by 10 stimuli:
- ***puls1***: train of pulses 100Hz, 5kHz LowPass
- ***ramp1***: sawtooth 100Hz , 5kHz LP
- ***triangle1***: triangle 100Hz, 5kHz LP
- ***breathy11***: sine wave 100Hz + 2 harmonics, amplitude ratio 16:4:1 from lowest to highest frequency
- ***breathy21***: *breathy11* + band noise 1-5kHz, noise amplitude 1/16 of the amplitude of the fundamental frequency sine wave
- ***whisper1***: *breathy11* + band noise 1-5kHz, noise amplitude similar to the amplitude of the fundamental frequency sine wave
- ***creak1***: train of pulse 100Hz, 5kHz LP , single peaks set to zero (irregular)
- ***falsetto11***: train of pulses 400Hz, 5kHz LP
- ***falsetto21***: train of pulses 400Hz, 5kHz LP + band noise 1-5kHz, noise amplitude similar to the amplitude of the fundam. freque. sine wave
- ***falsetto31***: train of pulses 400Hz, 5kHz TP, single values set to zero (irregular) + band noise 1-5kHz, noise amplitude similar to the amplitude of the fundam. frequ. sine wave

A "female voice" was designed by changing the fundamental frequency to 200 Hz.

In a listening test 45 pairs for each ("male/female") voice were presented to 10 naive listeners. The duration of each stimulus was 3 sec separated by a pause of 1 sec between the stimuli and 3 sec between the pairs. The listeners judged the similarity of the stimuli within each pair on a linear rating scale of 6 cm in length.
For each stimulus pair the average similarity and standard deviation was calculated and the pairs were arranged in an order of decreasing similarity.

Results

Pairs with differences in **fundamental frequency** (factor 4) are perceived as most different.
After that the next strongest differences were perceived in combinations of either harmonic or turbulent signals with **irregular signals**.
The difference of the "normal" irregular signal (*creak*) and the irregular signal with higher fundamental frequency (*falsetto3*) was judged less strong, even though the difference in the fundamental frequency itself would lead to a strong perceived difference. This shows that the irregularity of a voiced signal is perceived as a specific sound quality independent of the fundamental frequency of the signal.
Combinations of harmonic and **turbulent signals** are perceived less different than combinations with irregular signals.
The difference between the "normal" turbulent signal (*whisper*) and the turbulent signal with higher fundamental frequency (*falsetto2*) was judged less strong, even though the difference in the fundamental frequency itself would lead to a strong perceptual difference: Noisy components are also perceived as a specific sound quality independent of the fundamental frequency of the signal.
The perceptual difference of signals that differ strongly in the **spectral damping** of their

harmonics (*puls/breath1*) is placed in the same region of the rating scale as the perceived difference of combinations of harmonic signals and signals with strong turbulent components. This shows that the difference in the spectral damping of two signals is as important for the perceived sound quality as the criterion harmonic (voiced) versus turbulent (voiceless).

There were no systematic differences in the results for the "male" and "female" voice.

Experiment B - Part 2

The synthetic stimuli for both ("male/female") voices were filtered with the formants from a completely voiced utterance to achieve a more speechlike quality of the stimuli. The filtering can be interpreted as a spectral damping of some frequency regions, thus the stimuli are more equal in the physical domain after the filtering, but there is no "physical" reason why the order of perceived differences should change. The design of the listening test was similar to that used in the first part of Experiment B.

Results

In general the pairs of speechlike stimuli were judged as being more similar to each other than the synthetic stimuli without formant structure.
Combinations of "normal" frequency and *falsetto* were judged as much more different than in the first part of the experiment.
Differences between noisy, irregular and "normal" *falsetto* signals were judged las being ess than in the first part of the experiment.

The listening test shows, that the result from psychoacoustic experiments with "meaningless" sounds are not necessarily valid for speech perception. In speech perception an acoustic parameter can have a different effect than it has for the perception of a "meaningless" sound.

In general, most results from Part 1 of Experiment B are also valid for speechlike stimuli.

4. SUMMARY

Historically the acoustic parameter harmonic versus turbulent signal components was important in speech analysis and synthesis, because it represents the phonetically relevant feature voiced vs. voiceless which differentiates between phonemes.
The perception experiments have shown that irregularities (for example in creaky voice) and spectral damping (maximum spectral damping in lax or soft voice, minimum spectral damping in tense or shouted voice) lead to similar or even stronger differences in sound perception and should therefore receive more attention in speech analysis, recognition and synthesis.

5. REFERENCES

[1] Klasmeyer, Gudrun (1996) "Perceptual Cues for Emotional Speech", Proceedings for the ESCA Workshop "The Auditory Basis of Speech Perception", Keele

[2] Gobl, Christer (1989) "A preliminary study of acoustic voice quality correlates", STL-QPSR 2-3

[3] Klasmeyer, Gudrun & Sendlmeier, Walter F. (1995) "Objective voice parameters to characterize the emotional content in speech", Proceedings ICPhS 95, Stockholm, Vol. 1, pp. 181-185

[4] Laver, John (1994) "Principles of Phonetics", Cambridge University Press

[5] Michaelis, Dirk & Strube, Hans W. (1995) "Orthogonale akustische Stimmgüteparameter zur Stimmtherapiedokumentation", in Fortschritte der Akustik, DAGA95, Saarbrücken, Band 2, S.1035-1039