Voice Quality and Dialogue Structure

Kathrin Claßen & Wolfgang Wokurek
Institute of Natural Language Processing, University of Stuttgart, Germany
E-mail: {kathrin.classen, wolfgang.wokurek}@ims.uni-stuttgart.de

ABSTRACT

The primary aim of this acoustic study was to investigate spectral properties of first and repeatedly mentioned word tokens in task-oriented dialogues. Secondary aims were to assess the influence of word position in the phrase, role of speaker in the dialogue, and familiarity between participants, on these properties. The acoustic measurements comprised automatic spectral analysis of the glottal parameters open quotient (OQ), glottal opening (GO), skewness of glottal pulse (SK), rate of closure (RC), amplitude of voicing (AV), and completeness of closure (CC). Results of our Map Task experiments indicate significant spectral differences in the glottal parameters OQ, SK, RC, AV for the variables familiarity of participants and role of the speaker in the dialogues. First versus repeatedly mentioned word tokens differ significantly only in OQ and RC. Furthermore, no effects were found for the variable word position in the phrase.

1. INTRODUCTION

Previously, we have shown that voice quality is a domain of word stress correlate in German [1]. Based on studies by Sluijter et al. [2] we investigated spectral properties in order to infer laryngeal and sublaryngeal activity during the production of word stress. Results indicate that spectral tilt is significantly steeper in unstressed than in stressed syllables. Emphasis lies on the importance of measuring the amplitude correlate of word stress not as overall intensity, but with respect to the balance between amplitudes in the low-frequency and the mid-to-high frequency domain of the spectrum.

For Swedish, Heldner [3] demonstrated that spectral emphasis is a reliable acoustic correlate of focal accent. In these studies statistically significant differences between focally and non-focally accented words in any position of the phrase could be demonstrated.

Aims of the study

- The primary aim of the present study was to investigate spectral properties of first and repeatedly mentioned word tokens in spontaneous dialogues.
- Secondary aims of this study were to assess the influence of word position in the phrase, role of speaker within the dialogue, and familiarity between participants, on these properties.

2. MATERIAL AND METHODS

2.1 Subjects

16 native speakers (8 married couples) of German (Baden-Württemberg), age 57 to 75 years, performed task-oriented dialogues.

2.2 Speech material

Based on the established English *HCRC Map Task* [4], as well as an Italian version [5] and one German map [6], a German Map Task [7] was developed.

Task-oriented dialogues were carried out under the following conditions:

- Two participants of the experiments had slightly different versions of a simple map with approximately 10 landmarks on it *(discrepancies in placement and positioning of landmarks).*
- In one of the maps a route connecting the landmarks was printed. This participant of the experiments was intended to be the instruction giver of the dialogue (*Instruction Giver*, IG).
- The task of the second participant in the dialogue (*Instruction Follower,* IF) was to replicate the route.

Both participants of the dialogue were unaware as to the purpose of the experiment. They had no eye contact. The dialogues were repeated with a familiar (husband/wife) and an unfamiliar partner twice as *Instruction Giver* (IG), twice as *Instruction Follower* (IF).

2.3 Recording procedure

Recordings were taken in a sound-proof room at the Phonetics Department of the Institute of Natural Language Processing (IMS), University of Stuttgart, digitized at a sampling rate of 16 kHz/16 bit. Data analysis was carried out on an SGI work station using the signal processing package ESPS/xwaves (Entropic Inc.) [12].

Pre-analysis and inclusion criteria

After orthographic transcription of the dialogues, a conversational move analysis following Kowtko et al. [8]

 ISBN 1-876346-48-5

was carried out. *Instruct* and *explain moves* were selected for acoustic analysis, comprising 606 first and repeatedly mentioned landmark names in different positions in the phrase *(inphrase/phraseend)*.

2.4 Acoustic measurements

Acoustic measurements were carried out on stressed vowels of the landmark items. They included long and short vowels of [i, a, u, e, and o]-vowel types.

The automatic signal analysis procedure is based on amplitude and frequency measurements at harmonic spectral peaks. It is described in more detail in [11]. The spectral peaks are searched starting at integer multiples of the fundamental frequency (F0). The F0 estimate is computed by a correlation function based pitch tracking algorithm. Furthermore, harmonic peak amplitudes next to the formants are required. Formant frequencies F1-F4 and bandwidths B1-B4 are LPC based estimates. We use the ESPS program [12] **formant** to estimate these parameters.

The spectral peak amplitude search requires a narrow band spectrum with narrowly spaced spectral samples. This is done by the Fast Fourier Transform (FFT) using a Hamming window of 25ms duration and an FFT order of 14. This window duration yields a narrowband analysis. Independently of the window duration, the FFT order determines the number of frequency samples to be 2^{14}=16384 and the spectral samples are spaced apart by about 1Hz for our 16kHz sampled signals.
To achieve vowel independent glottal parameter estimates we adapt the notion of [2], [9] and [1] and reduce the vowel influence by removing the contribution of the resonances to the spectral amplitude. The spectral gain due to each formant is estimated by the model developed in [10]. Each formant model is specified by two parameters, the frequency and the bandwidth of the formant. As usual, the amplitude measurements are made in decibels and so the formant influence is removed by subtracting the resonance amplitude from the spectral amplitude reading.

If the formant gain is estimated at frequencies several bandwidths apart from the center frequency, a simplified model equation may be used. The simplified model assumes a zero bandwidth and has only the formant frequency as a parameter. In the current developement phase of the voice parameter estimators both types of models are employed to see which produces the better estimates.

Open quotient

```
OQ = H1* - H2*; H1* = H1P - dH1; H2* = H2P - dH2
dH1 = Decibel (( F1 **2 +( B1 /2)**2)/
      sqrt((( F0P - F1 )**2+( B1 /2)**2)
          *(( F0P + F1 )**2+( B1 /2)**2))))
Decibel = 20/log(10)*log
```

The influence of the first formant is removed from the amplitudes of the first two harmonics resulting in H1* and H2*. Their difference on a logarithmic scale (in decibels) serves as the estimate of the open quotient.

Glottal opening `GO = H1* - A1P`

Including the bandwidth in the formant model improves the estimate for signals with high F0 or low F1.

Skewness

```
SK = H1* - A2*; A2* = A2P - dA21 - dA23 - dA24
dA21 = Decibel ( F1 **2/abs( F1 **2- F2P **2))
dA23 = Decibel ( F3 **2/abs( F3 **2- F2P **2))
dA24 = Decibel ( F4 **2/abs( F4 **2- F2P **2))
```

Removing the influence of F1, F3, and F4 on F2 improves the skewness estimate.

```
SK' = SK / Octaves02; log2 = 1/log(2)*log
Octaves02 = ( log2 ( F2P ) - log2 ( F0P ))
```

This spectral slope measured in decibels per octave improves the skewness estimate.

Rate of closure

```
RC = H1* - A3*; A3* = A3P - dA31 - dA32 - dA34
```

Removing the influence of F1, F2, and F4 on F3 improves the rate of closure estimate.

```
RC' = RC / Octaves03
Octaves03 = ( log2 ( F3P ) - log2 ( F0P ))
```

This rate of closure measured in decibels per octave improves the skewness estimate.

Amplitude of voicing `AV = H1*`

Completeness of closure `CC = B1`

Uncertainties

LPC based formant estimates are not equally reliable for all types of speech. The fundamental frequency and the spectral slope of the source spectrum are important influential quantities. This is due to the fact that LPC places the poles in order to account for as much signal energy as possible.

Unfortunately LPC formant estimates of female speech are more likely to fail compared to that of male speech. This may be due to the higher fundamental frequency forcing the harmonics to attract the formant estimates. Secondly incomplete glottal closure may result in a steeper spectral decay of the voiced excitation [13].

Window duration and the slope of the preemphasis filter are effective parameters of the LPC analysis. The baseline analysis is done using the standard parameters that correspond to a Hanning window of 25ms duration and a preemphasis of 0.7 which inludes low frequency signal

ISBN 1-876346-48-5

components and has a moderate preemphasis of the high frequency parts of the signal. It turned out that about 10% of the signals are put to the wrong place in the vowel triangle. However, an increased preemphasis parameter puts many of them to the right location. Our analysis procedure includes multiple analyses with the preemphasis parameters 0.7, 0.8, 0.9, 0.99 and 1.0, which increasingly put more weight on the high frequency signal components.

About 2% of the signals lead to spectral measurements that cause numerical error conditions when the voice source parameters are calculated. Currently these signals are considered as outliers and are excluded from the analysis.

2.5 Statistical analysis

Mean values for the glottal parameters open quotient (OQ), glottal opening (GO), skewness of glottal pulse (SK), rate of closure (RC), amplitude of voicing (AV), and completeness of closure (CC) as well as standard deviations were calculated for all 16 speakers and all vowel types. The Mann-Whitney-U-Test was used to analyse differences on these properties between (1) first and repeatedly mentioned word tokens, (2) influence of word position in the phrase, (3) role of speaker in the dialogue, and (4) familiarity between participants.

3. RESULTS

For the presentation of the results and the statistical analysis the data of all subjects (n=16) were either pooled together, or were analyzed by gender (female speakers, n=8; male speakers, n=8). In the t-tests first and repeatedly mentioned word tokens, influence of word position in the phrase, role of speaker in the dialogue, and familiarity between participants, were chosen as independent variables. The tests were run on the glottal parameters open quotient (OQ), glottal opening (GO), skewness of glottal pulse (SK), rate of closure (RC), amplitude of voicing (AV), and completeness of closure (CC). Those parameters that were found to be significant (probability of t-statistic $p<0.05$) are given in bold typed in tables 1-4).

3.1 First and repeatedly mentioned word tokens

	first mentioned	repeatedly mentioned
AV	40.11 (+/-9.31)	39.52 (+/-8.68)
OQ	**-5.82 (+/-10.64)**	**-8.79 (+/-12.18)**
GO	-3.63 (+/-12.52)	-4.20 (+/-11.13)
SK	11.04 (+/-5.09)	10.11 (+/-5.48)
RC	**15.17 (+/-5.1)**	**16.41 (+/-5.20)**
CC	123.74 (+/-118.04)	132.04 (+/-112.16)

Table 1: Mean values +/-standard deviations for the different spectral parameters (vertically) and independent variable (horizontally), first vs. repeatedly mentioned item, pooled across all speakers. Values are expressed in decibels (dB).

First versus repeatedly mentioned word tokens differ significantly in open quotient (OQ) and rate of closure (RC).

3.2 Word position in the phrase

No significant effects were found in any spectral parameter for the variable word position in the phrase.

	Inphrase	Phrase end
AV	39.25 (+/-9.53)	0.18 (+/-8.11)
OQ	-8.46 (+/-11.87)	-7.24 (+/-11.71)
GO	-4.59 (+/-12.33)	-3.41 (+/-10.68)
SK	10.08 (+/-5.26)	10.73 (+/-5.48)
RC	16.26 (+/-5.30)	15.79 (+/-5.12)
CC	133.18 (+/-115.78)	125.52 (+/-112.07)

Table 2: Mean values +/-standard deviations (in dB) for the spectral parameters and independent variable word position in the phrase *(inphrase/phrase end).*

3.3 Role of speaker

Significant differences in spectral parameters SK, RC and AV were found for role of the speaker in the dialogue.

	IG	IF
AV	**40.07 (+/-8.68)**	**35.47 (+/-9.97)**
OQ	-8.04 (+/-11.91)	-5.89 (+/-10.25)
GO	-3.82 (+/-11.64)	-6.35 (+/-10.58)
SK	**10.17 (+/-5.27)**	**12.94 (+/-5.95)**
RC	15.85 (+/-5.15)	18.06 (+/-5.60)
CC	131.24 (+/-116.14)	109.54 (+/-83.90)

Table 3: Mean values +/-standard deviations (in dB) for the spectral parameters and independent variable role of speaker *(Instruction Giver/Instruction Follower).*

3.4 Familiarity between participants

	known	unknown
AV	**40.23 (+/-8.53)**	**38.86(+/-9.35)**
OQ	**-9.33 (+/-12.50)**	**-5.55 (+/-10.18)**
GO	-3.67 (+/-11.38)	-4.58 (+/-11.87)
SK	**11.4 (+/-4.20)**	**8.64 (+/-6.47)**
RC	**16.31 (+/-5.41)**	**15.58 (+/-4.86)**
CC	123.90 (+/-112.13)	128.77 (+/-138.72)

Table 4: Mean values +/-standard deviations (in dB) for the spectral parameters and independent variable role of speaker *(Instruction Giver/Instruction Follower).*

ISBN 1-876346-48-5

Familiarity between participants *(known/unknown)* turned out to show the largest spectral differences. Significant differences in OQ, SK, RC and AV were found for all speakers.

3.5 Influence of gender on experimental results

When analyses 3.1-3.4 were separately repeated for male and female subjects, respectively, striking differences were found for each gender as compared to the pooled analysis presented above. For male speakers, familiarity between speakers, role of speakers and word position in the phrase were all significantly different for CC, while no such differences were observed for female speakers. In the pooled analysis CC did not show significant differences for any independent variable. Likewise, for male speakers AV and OQ showed significant differences for each condition except for word position in the phrase and role of the speaker, respectively, while for female speakers no such differences could be detected for any condition. In the pooled analysis, however, significant differences in AV and OQ became evident as outlined in tables 1-4.

4. CONCLUSIONS

All results of our Map Task experiments indicate significant differences in the glottal parameters AV, SK, RC, CC for the variables familiarity of participants (known/unknown) and role of the speaker (IG/IF) in the dialogues. No effects except for OQ were found for the variables first versus later mentioned word tokens and word position in the phrase: OQ differs significantly between first and later mentioned words. In conclusion, our experiments indicate that contrary to initial expectations there was no spectral difference in first and second word tokens. However, familiarity and role of the speaker (IF/IG) in the dialogue significantly affected spectral properties.

The independence of the voice quality parameter estimates from the vowel quality has been found to be improved by (i) compensating for the influence of two neighboring formants instead of one, (ii) predicting the formant influence including the bandwidth estimate, and (iii) measuring the spectral slope with logarithmic amplitude and frequency axes in decibels per octave.

REFERENCES

[1] K. Claßen, G. Dogil, M. Jessen, K. Marasek, and W. Wokurek, “Stimmqualität und Wortbetonung im Deutschen, “ *Linguistische Berichte,* vol. 174, pp. 202-245, 1998.

[2] A.M.C. Sluijter, S. Shattuck-Hufnagel, K.N. Stevens, and V. J. van Heuven, “Supralaryngeal resonance and glottal pulse shape as correlates of stress and accents in English,” *Proceedings of the International Congress of Phonetic Sciences 13 (Stockholm)*, vol 2, pp. 630-633, 1995.

[3] M. Heldner, “On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish,” *Journal of Phonetics,* vol. 31 (1), pp. 39-62, 2003.

[4] A. H. Anderson, M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, K., J. Kowtko, J. McAllister, J. Miller, C. Sottilo, H. Thompson, and R. Weinert, “The HCRC Map Task Corpus,” *Language and Speech,* vol. 34 (4), pp. 351-366, 1991.

[5] M. Grice and M. Savino, “Intonation and communicative function in a regional variety of Italian,” *Phonus,* vol. 1, pp. 19-32, 1995.

[6] M. Grice and R. Benzmüller, “Transcriptions of German intonation using ToBi-Tones - The Saarbrücken System,” *Phonus,* vol. 1, pp. 33-51, 1995.

[7] K. Claßen, “Map Task – Eine Version für das Deutsche,” *AIMS,* vol. 6 (4), pp. 65-83, 2000.

[8] J. Kowtko, S. Isard, and G. Docherty-Sneddon, “Conversational games within dialogue,” *HCRC/RP-31*, Edinburgh, 1993.

[9] K. M. Stevens and H. M. Hanson, "Classification of glottal vibration from acoustic measurements", in *Vocal Fold Physiology*, O. Fujimura and M. Hirano, Eds., pp. 147-170. Cambridge MA: Hiltop University Press, 1998.

[10] G. Fant, *Acoustic theory of speech production*, The Hague: Mouton, 1970.

[11] W. Wokurek and M. Pützer, "Automated corpus based spectral measurement of voice quality parameters", *Proceedings of the International Congress of Phonetic Sciences 15 (Barcelona)*, 2003.

[12] ESPS Programs Version 5.3.1, Entropic Inc., 1999.

[13] H. M. Hanson, E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data", *Journal of the Acoustic Society of America*, 106: pp. 1064-1077, 1999.