Automated Corpus Based Spectral Measurement of Voice Quality Parameters

Wolfgang Wokurek[*] and **Manfred Pützer**[†]

[*] Institute of Natural Language Processing, University Stuttgart, Germany
wokurek@ims.uni-stuttgart.de

[†] Institute of Phonetics, University of the Saarland, Saarbrücken, Germany
puetzer@coli.uni-sb.de

ABSTRACT

An automated procedure for the estimation of voice quality parameters for whole corpora of segmented speech recordings is presented. It is used to evaluate voice quality parameters of speech segments too numerous for manual measurements and to improve the estimation formulas. The analysis procedure is tested on a subset of 120 sustained vowels from a database of recordings of normal-voice and pathological-voice speakers. The voice quality parameters are investigated for significant differences between normal and pathologic speech for both male and female speakers.

1 INTRODUCTION

Spectral estimates that correlate with the voice quality 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) have been identified by [1], [2], and [3]. The common structure of these estimators are ratios of spectral peak amplitudes of the harmonics. Usually the amplitude measurements are done on a logarithmic scale (decibels) where the amplitude ratios appear as differences. For instance, a measure for the open quotient is the difference between the first two harmonic peaks with the formant influence removed. H1 and H2 are defined as the peak amplitudes (in dB) at the fundamental frequency F0 and its doubled value, 2×F0. Due to resonance, H1 and H2 benefit from a low first formant. Hence, assuming the same voiced excitation, H1 and H2 are vowel dependent. Using a model for the amplitude response of the first formant [4] this vowel dependency may be removed (subtracted on the dB scale), resulting in H1* and H2*. Their difference OQ = H1* - H2* serves as a parameter related to the open quotient.

2 SIGNAL ANALYSIS

Sustained vowels were the objective of this study. These vowels were recorded and stored each in a separate file. The subsequent signal analysis requires a minimum duration of two frames (35ms = 25ms window duration + 10ms step width) which was met by instructing the subjects to utter each vowel for more than a second.

2.1 Frequency and amplitude measurements

The spectral correlates of glottal parameters involve peak amplitudes next to formants. Due to the harmonic structure of vowels such peak amplitudes are expected near integer multiples of the fundamental frequency. Hence, three different types of measurements are required: (i) the fundamental frequency F0, (ii) the formant frequencies F1, F2, ... together with their bandwidths B1, B2, ... and (iii) the FFT spectrum for amplitude measurements. We employ ESPS programs [5] to calculate (i) - (iii) and a PERL script for rapid prototyping of the implementation.

In particular, the ESPS program **formant** solves (i) and (ii). It uses a correlation function based pitch tracking algorithm [6] to estimate F0 and linear prediction to estimate formant frequencies and bandwidths. The program **fft** uses the Fast Fourier Transform to calculate spectral amplitudes (iii). A Hamming window of 25ms duration is used in order to include at least two pitch periods even for male speech. This is necessary for the amplitude spectrum to contain a first glance of the harmonic line structure. Clearly the harmonic lines become more marked at higher pitch. Using this window results in a frequency resolution comparable to a narrowband spectrogram. For the automatic peak search many frequency samples are required. This is achieved by using an FFT order of 14, resulting in $2^{14} = 16384$ spectral samples that are spaced apart by about 3Hz for our 50kHz sampled signals. Figure 1 shows such an amplitude spectrum taken from the center of an [a:] sound. Rounding the frequency readings to Hz is granular enough, even in

ISBN 1-876346-48-5

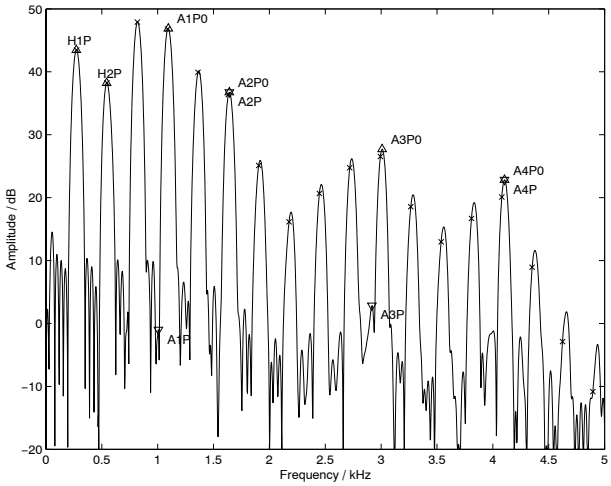

Figure 1: Narrowband spectrum of a single frame with peak search based amplitude readings. The x-marks show the integer multiples of the somewhat too low F0 estimate.

spite of the coarse underlying LPC model.

```
# 1000FNan_SND.sd
fem=1 mal=0 nor=1 pat=0 a=1 i=0 u=0
t=     0.0245  F0  = 244.3  F1  = 982.  F2  = 1502
               2F0  = 488.7  B1  = 120.  B2  = 123.
                             F10 = 977.  F20 = 1466
                             F1P = 1000  F2P = 1501
  H1  = 40.37  H2  = 37.96  A1P = 51.9  A2P = 45.6
  F0P = 250.2  2F0P= 500.4  F1P0= 1000  F2P0= 1501
  H1P = 40.53  H2P = 38.62  A1P0= 51.9  A2P0= 45.6
t=     0.0345  F0  = 248.0  F1  = 984.  F2  = 1509
               2F0  = 496.0  B1  = 127.  B2  = 140.
                             F10 = 992.  F20 = 1488
                             F1P = 1007  F2P = 1510
  H1  = 41.08  H2  = 38.19  A1P = 51.9  A2P = 44.2
  F0P = 253.2  2F0P= 503.5  F1P0= 1007  F2P0= 1510
  H1P = 41.30  H2P = 38.31  A1P0= 51.9  A2P0= 44.2
```

Table 1: Spectral measurement protocol for the first two frames of the signal `1000FNan_SND.sd` including classification attributes. The measurements for the third and fourth formant are not shown.

A sample of the automatic measurement protocol is shown in table 1. After the line showing the signal's file name follows a line with the attributes of that recording. Then a sequence of blocks of framewise measurements follows. The format is fixed and designed to keep the measurement protocol readable not only for machines. Each block starts with its epoch in seconds. The first two lines of each block show the results of the F0 and formant measurements (i), (ii). 2F0 is simply the double of F0 and represents the frequency of the second harmonic. F1 - F4 are the LPC estimates of the first four formants and B1 - B4 those of their bandwidths.

The third line shows the frequency of the harmonic next to each formant estimate. F10 = 977 Hz is the fourth harmonic and no other harmonic is closer to F1 = 982. If any of the formant estimates is very close to a harmonic (Fi = Fi0) it could well be that LPC was influenced by the harmonic structure of the signal instead of showing the resonances of the vocal tract.

The fourth line of each frame shows the spectral peaks closest to each of the formant estimates. These frequencies are F1P - F4P and the corresponding spectral amplitudes A1P - A4P are shown in line five together with the amplitudes at the fundamental frequency estimate of the first line (H1 at F0 and H2 at 2F0).

The final two lines (fifth and sixth) focus further on the harmonic structure. They contain improved harmonic peak measurements at the integer multiples of F0.

2.2 Exclusion of outliers

Signals were excluded due to one of three reasons: (i) shortness yielding less than two frames (ii) formant estimates misplacing the sound in the vowel triangle and (iii) spectral measurements producing numerical errors in computing one of the glottal parameter estimates.

Single frame analyses were discarded mainly as a precaution not to have unreliable samples in this first study based on automated analysis. Even though the unreliability of such data was only suspected, we wantet to compute the standard deviation of each sound measurements and it takes a minimum sample size of two to do so.

The formants are estimated as the pole frequencies of the linear prediction analysis. A standard preemphasis of high frequencies controlled by a parameter of 0.7 puts most vowels in the right place in the vowel triangle in about nine of ten cases. However, the rest ended up at the wrong vowel quality. First, the vowel quality of the recording was checked acoustically and verified. Then the preemphasis parameter was increased to 0.8, 0.9, 0.99 and 1.0. It turned out that about two thirds of the misplaced vowels changed to the right place when analysed with one of these preemphasis parameters. The remaining 3% of the signal set were excluded. Shortening the analysis window of 25ms might improve the analysis of high pitched signals.

The evaluation of formulas that include divison by a difference is not possible if the difference vanishes (error condition: division by zero). It turned out that some frames in about 1% of the analysed speech recordings suffer from this fact. All attempts to predict this error condition from the spectral measurements in advance without full evaluation of the estimation formulas failed in the sense that many good files were marked for exclusion and some of the non-marked were bad. Only full evaluation of all spectral measurements with all formulas showed which signals to exclude.

ISBN 1-876346-48-5

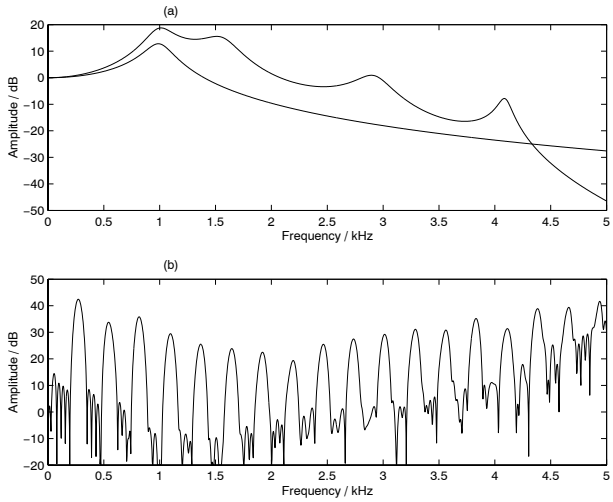

Figure 2: (a) Model for the amplitude influence of formants (b) Spectrum with formant model subtracted

2.3 Formant model

The influence of a certain formant with center frequency F and bandwidth B on the spectral amplitude measurement at frequency f is predicted by the frequency response [4, p.53 eq.1.3-5b]

$$A(f;F,B) = \frac{F^2 + (\frac{B}{2})^2}{\sqrt{(f-F)^2 + (\frac{B}{2})^2}\sqrt{(f+F)^2 + (\frac{B}{2})^2}} \quad . \tag{1}$$

According to the source-filter model, each formant comprises one such factor multiplied by the source spectrum $G(f)$. Together with the pick-up factor $R(f)$ containing mainly the radiation, the spectral amplitude of the recorded signal is

$$S_{dB}(f) = R_{dB}(f) + \sum_{n=1}^{4} A_{dB}(f;F_n,B_n) + G_{dB}(f) \tag{2}$$

with the amplitudes S, R, A, and G in decibels

$$< \text{amplitude} >_{dB} = 20\log_{10}(< \text{amplitude} >) \quad . \tag{3}$$

Figure 2a shows both the spectral amplitude model of the first formant and the combined model of all four formants (the $\sum$-term in eq.2). This formant influence may be removed by subtraction, as shown in figure 2b. According to eq.2 it equals $G_{dB}(f) + R_{dB}(f)$, the sum of the source spectrum and the radiation transfer function. The latter very likely causes the rising shape starting around 2.2kHz.

To remove, e.g., the influence of the first formant on the amplitude measurement H1 at F0, the subtraction $H1_{dB} - A_{dB}(F0;F1,B1)$ is appropriate.

2.4 Estimation of voice quality parameters

Voice quality parameters were applied in their known format ([1], [2], and [3]) and varied in a systematic way: no formant correction, formant correction neglecting the formant bandwidth, formant correction including the formant bandwith; correction of only the neighboring formants; correction of all other formants.

```
Open quotient: OQ = H1* - H2*; H1* = H1P - dH1;
   H2* = H2P - dH2; Decibel = 20/log(10)*log;
   dH1 = Decibel (( F1 **2 +( B1 /2)**2)/
         sqrt((( F0P - F1 )**2+( B1 /2)**2)
             *(( F0P + F1 )**2+( B1 /2)**2))))
Glottal opening: GO = H1* - A1P
Skewness: SK = H1* - A2*
   A2* = A2P - dA21 - dA23 - dA24
   dA21 = Decibel ( F1 **2/abs( F1 **2- F2P **2))
   SK' = SK / Octaves02; log2 = 1/log(2)*log
   Octaves02 = ( log2 ( F2P ) - log2 ( F0P ))
Rate of closure: RC = H1* - A3*
   A3* = A3P - dA31 - dA32 - dA34
   RC' = RC / Octaves03
   Octaves03 = ( log2 ( F3P ) - log2 ( F0P ))
Amplitude of voicing: AV = H1*
Completeness of closure: CC = B1
```

Table 2: Equations for parameter estimation.

Table 2 shows some of the formulas used, to demonstrate the current state of the interpreter. Interpretation was chosen to allow flexible changes to the parameter estimation formulas applying them repeatedly to the same spectral raw measurements as in table 1.

Subtracting the gain of the first formant from the amplitude measurement of the first harmonic H1P is expressed by the term H1* = H1P - dH1. The expression for dH1 implements eqns.1 and 3.

Skewness and rate of closure estimates so far are only amplitude ratios (decibel differences). Both peak frequencies depend on F0 and the higher one also on the formant and hence on the vowel quality. To reduce the vowel quality influence, spectral slopes SK' and RC' are defined which relate the vertical decibel distance to the horizontal number of octaves. This improvement is confirmed by statistical factor analysis where SK' and RC' load stronger to their factors than their competitors.

3 EVALUATION

The recordings of 40 German speakers were used for the present study. All speakers produced sustained vowels [i:, a:, u:] at normal pitch. There were two groups of speakers. The first group consisted of 10 male and 10 female pathological speakers with organic or functional voice disorders. The second group consisted of 10 male and 10 female normal speakers with no known speaking or hearing problems, matched in age to the pathological speakers. The microphone signal was recorded simultaneously with the EGG signal

ISBN 1-876346-48-5

in a sound-treated room, using a neckband condenser microphone (NEM 192.15, beyerdynamic). The signal was fed directly into a Computerised Speech Lab station (CSL; model 4300B) at a sampling rate of 50 kHz to reduce the temporal quantisation error to 0.02 ms and with a 16 bit amplitude resolution. Only the microphone signal was analysed for this study.

3.1 Statistics

The statistics were carried out using SPSS version 10.0. First, multi-variate analyses of variance were performed to test the parameters' ability to distinguish the two groups. This procedure was carried out separately for male and female speakers. Additionally, in a factor analysis, the relative weightings of the parameters within the factors were tested. Finally, a clustering technique was applied out on the basis of the parameters to differentiate subgroups of normal and pathological speakers for each gender.

3.2 Multi-variate analyses of variance

In a multi-variate analysis of variance, a significant differentiation of the two groups is achieved on the basis of the parameters for male and for female speakers.

female	*				*	
variable	OQ	GO	SK	RC	AV	CC
male	*	*	*		*	*

More parameters are valid to differentiate non-pathological and pathological male speakers at a significance level of 5%.

3.3 Factor analysis

For the two genders, factors are found which correspond to the physiological orientation of the analysis. The factors shown represent 90% of the variance observed.

factor	1	2	3	4	5
female	RC	SK	OQ,GO,CC AV	OQ,AV GO	
male	RC	SK	OQ,GO,CC	OQ,AV	AV

These groups of variables load more than 0.75 each.

3.4 Clustering technique

With the clustering technique a voice profile differentiation is achieved for normal male and female speakers as well as for pathological male and female speakers. Non-pathological female and male subjects are grouped into two clusters each. Six clusters are required for our female and three for our male pathological speakers.

3.5 Discussion

The differentiation of normal and pathological voices is dependent on gender-specific vocal properties. It is easier to distinguish male normal voices from male pathological voices. It is true that the differentiation is also achieved for female speakers, but there are less parameters which differentiate them. The reason for this lies perhaps in the less complete glottal closure for female speakers, leading to more energy loss at the glottis and a stronger spectral tilt [7]. Because pathological female voices cannot easily be distinguished from normal female voices, the parameterisation of female voices needs to be improved. The grouping of physiologically related parameters within one factor indicates the physiological validity of the factor analysis. The voice profile differentiation for normal male and female speakers and for pathological male and female speakers, achieved by the clustering technique underlines the validity of the parameters. The reliability of this potential needs to be tested on a larger number of signals.

4 CONCLUSION

Significant gender-differentiated distinction between normal and pathological voices was found. This encourages further investigations of the whole corpus by this technique. The compensation of the formant resonances is improved by (i) including the bandwidth estimate, (ii) removing more neighboring formants, and (iii) measuring the spectral slope with logarithmic amplitude and frequency axes in decibels per octave.

REFERENCES

[1] Kenneth M. Stevens and Helen M. Hanson, "Classification of glottal vibration from acoustic measurements," in *Vocal Fold Physiology*, Osamu Fujimura and Minoru Hirano, Eds., pp. 147–170. Cambridge MA: Hiltop University Press, 1998.

[2] Agaath Sluijter, *Phonetic Correlates of Stress and Accent*, The Hague: Holland Academic Graphics, 1995.

[3] Kathrin Claßen, Grzegorz Dogil, Michael Jessen, Krzysztof Marasek, and Wolfgang Wokurek, "Stimmqualität und Wortbetonung im Deutschen," *Linguistische Berichte*, vol. 174, pp. 202–245, 1991.

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

[5] "Esps Programs Version 5.3.1," *Entropic Inc.*, 1999.

[6] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '83*, vol. 8, pp. 1352–1355, 1983.

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

ISBN 1-876346-48-5