Vowel Dependence for Electroglottography and Audio Spectral Tilt

Laura Enflo
Dept. of Speech, Music and Hearing, Royal Institute of Technology (KTH)

Abstract

The spectral tilt has been calculated for the vowels from audio signals and from the derivative of electroglottography (EGG) signals in a Swedish one-speaker corpus of 5277 sentences. Vowel dependence has been found for the audio spectral tilt but not for the EGG spectral tilt, which also gets steeper as the sound pressure level increases. The EGG spectral tilt values had an average standard deviation of 1.3 dB/octave and the corresponding standard deviation value for the audio spectral tilt was 2 dB/octave.

Introduction

Spectral tilt, or spectral slope, is an important parameter in voice synthesis and voice perception. One way of defining it is ‘the least squares linear fit to the harmonic peaks in the spectrum of the glottal waveform calculated from an audio speech signal’. A less steep spectral tilt corresponds to a louder voice and when loudness decreases, the spectral slope has been confirmed to increase (Fant and Lin, 1988 & Hanson, 1997). Spectral tilt has also been linked to different voice qualities. For example, steeper spectral tilt values have been found for female voices perceived as breathy by speech therapists (Karlsson, 1988 & 1992). In automatic detection of prominence (emphasis) in speech, on the other hand, the spectral tilt parameter has been found to be a weak indicator (Al Moubayed et. al., 2010).

The spectral tilt is, in consequence, dependent on the placement of formant peaks. Formants are known to provide the necessary acoustic information for vowel identification. For each vowel type, however, the formant frequencies have a high variability across speakers and phonetic contexts (e.g. Hillenbrand et. al., 1995). Independent manipulations of the formant frequency and spectral tilt parameters have been carried out in a more recent experiment. Each of the parameters had on their own a strong impact on the perception of vowel quality. Nevertheless, the perceptual effect of changes in spectral tilt was lessened in stimuli in which formant frequencies, as in natural-sounding speech, changed over time. Therefore, spectral tilt might be of little relevance to the process of identifying vowels correctly (Kiefte and Kluender, 2005).

Electroglottography, henceforth EGG, is a widely used technique for the assessment of vocal-fold contact during phonation (e.g. Orlikoff, 1998). An electroglottograph is provided with two electrode plates, which are placed on each side of the larynx. The idea with EGG is to send an electrical current from one electrode to the other and record the amplitude of this signal. When the vocal folds are closed, the signal can pass, such that the current is higher, and when the glottis is open, the current is lower. Several studies have shown that EGG is a more robust technique than audio for fundamental frequency (f0) determination thanks to its simpler waveform (e.g. Vieira et. al., 1996).

Many research projects have aimed to acquire information about the glottal voice source. A widely popular model is the Fant theory of a separated voice source and filter (Fant, 1960). In Fant’s model, the source can be modeled with the derivative of the glottal flow volume velocity. Before glottal source parameters can be obtained from the audio speech signal, however, vocal tract resonances need to be removed with the help of inverse filtering (Miller, 1959 & Rothenberg, 1973). Although the inverse filtering procedure can be done digitally, it often needs time-consuming modifications afterwards which make it harder to obtain the amount of data needed for a satisfactory statistical evaluation. Therefore, a considerable amount of research has been carried out on how to use the speech spectrum instead of time-domain estimations. One discovery has been that H1*-A3* (the amplitude of the third formant relative to that of the first harmonic) is correlated with the source spectral tilt, except when the first harmonic is weak (Hanson, 1997). A recent study has shown that H1*-A3* is vowel dependent (Iseli et. al., 2007).

Although the EGG signal is not a measure of glottal flow, it can give information about the glottal voice source without the use of the audio signal, which according to Fant's theory is strongly influenced by the linear filter that can be modeled from the vocal tract. In brief, audio spectral tilt is determined by both the source and the filter, whereas the EGG spectral tilt is affected only by the source. If EGG spectral tilt shall be used in speech synthesis and analysis in the future, it would be valuable to know whether it is vowel dependent or not. Since vowel identities are mainly determined by the shape of the vocal tract, it is likely that the EGG has no impact on this matter. However, since the result in the study by Iseli et. al. (2007) is contradictory to this idea, it is worth while to make an investigation.

MF Corpus

The MF corpus consists of 5277 sentences from newspaper texts and literature read by a professional male Swedish actor and was recorded in 2002 in a studio. The sampling frequency of the data is 16 kHz. Each audio signal has a corresponding signal with the EGG derivative.

Annotation was made with Sjölander's speech aligner (Sjölander, 2003). Occasional problems in the annotation of unusual words or names resulted in exclusion of some sentences, so that the speech material in this study added up to 5114 sentences in total. From this set of data all vowels with duration of at least 31 ms were picked (i.e. 500 data points), with the consequence that less than 7 % of all the vowels were discarded for the purpose of obtaining more reliable data. The average duration of a vowel was 117 ms.

The following vowels are represented in the corpus: [a, ɑ:, e, e:, ɪ, i:, u, u:, ɵ, ʉ:, ʏ, y:, ɔ, o:, ɛ, ɛ:, æ, æ:, ø, ø:, œ, œ:]. The total number of picked vowels was 35990 from the EGG files and 36585 from the audio files, with the distribution between vowel types as seen in Table 1.

Table 1. Number of picked vowels from the MF corpus, sorted in order of frequency when calculating the EGG spectral tilt (EGG ST). Number of vowels picked for calculating the audio spectral tilt (Audio ST) are also included.

Vowel	**EGG ST**	Audio ST
a	**3866**	3945
ɪ	**3269**	3346
ɔ	**2634**	2665
ɛ	**2633**	2749
e	**2616**	2693
ɑ :	**2594**	2649
e:	**2490**	2514
o:	**2148**	2066
i:	**2003**	2059
ʉ :	**1628**	1631
u:	**1593**	1554
u	**1436**	1448
ɵ	**1374**	1401
æ	**1038**	1089
ø:	**886**	911
œ	**836**	827
ʏ	**678**	698
y:	**597**	594
ɛ :	**541**	555
œ:	**443**	473
ø	**376**	394
æ:	**311**	324
Total	**35990**	36585

Parameters

For each of the vowels, the respective sound pressure level, loudness, duration, audio spectral tilt and EGG spectral tilt were calculated. A more thorough description of these features follows below:

Sound pressure level: the logarithmic value of the sound pressure of the speech signal relative to the human hearing threshold p_{ref}=20e-6 Pa (Timoney et. al., 2004).

Loudness was calculated according to the standard ITU-R BS 1770-1 (ITU Original) (ITU-R, 2006 & Nygren, 2009).

Duration: The start time was subtracted from the end time for each respective vowel and the results were displayed in seconds.

Audio spectral tilt and *EGG spectral tilt*: The spectral tilt values were obtained in a similar way for audio and for EGG. Three methods were tried out and evaluated on the EGG files. The best method was then chosen for both audio and EGG spectral tilt calculations. In the first method investigated, all separate spectrum peaks were found (using the findpeaks function by O'Haver, 1995) and a linear line which best fitted the data was calculated in a least-squares sense. The slope of the line provided the spectral tilt value. For the second method explored, the calculations were made in the same way, but with a cut-off frequency threshold corresponding to 6000 Hz on a logarithmic scale, in order to erase the impact of the peaks in the frequency range above the threshold, which was thought to be a source of error. Although the median values from both of these two methods showed stability (with standard deviations of 0.13 versus 0.15, respectively), the standard deviation of the average values was 40% higher for the first method in comparison to that of the second method. In addition, the average values differed greatly from the median values (on average 3.8 dB/octave) in the first method. In the second method, this difference was only 0.4 dB/octave on average. In the third method the calculations were made in the same way as for the first two methods, but with a cut-off frequency threshold corresponding to 3000 Hz on a log scale. In addition, spectral tilt data suffering from one of the three following problems were elimited from the analysis: 1) measured point was singular, 2) value was measured on too high frequencies, at least corresponding to 200 Hz, which is about twice as high as the average frequency for the speaker's voice, 3) signal was noisy in the middle frequency range, resulting in positive values. 10 % of the data had to be discarded due to these three problems.

The average value per vowel had a standard deviation of 10.5 dB/octave in the second method. For the third method, the corresponding standard deviation was 1.3 dB/octave. Consequently, the third method was used in this experiment.

Analysis

All of the parameters mentioned above were implemented in and the analysis was performed with Matlab 7.3.0 (release 2006b). The calculations were made on the 500 middle data points of each vowel, except for the vowel duration parameter. Statistical analysis was carried out in SPSS.

Results

The sound pressure level is 77 dB ± 3.4 dB for the audio spectral tilt values (Figure 1).

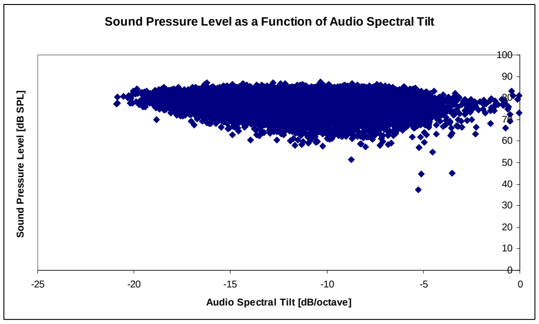

Figure 1. Sound pressure level as a function of audio spectral tilt.

For the EGG spectral tilt, there is a noticeable decrease of the sound pressure level as the spectral slope is decreased (Figure 2).

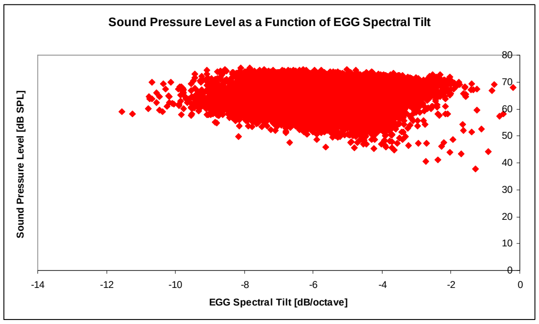

Figure 2. Sound pressure level as a function of EGG spectral tilt.

Statistically significant differences for the audio spectral tilt values were found between most vowels. The median values (Figure 3) have the same values on standard deviations as the average audio spectral tilt (Figure 4); 2 dB/octave.

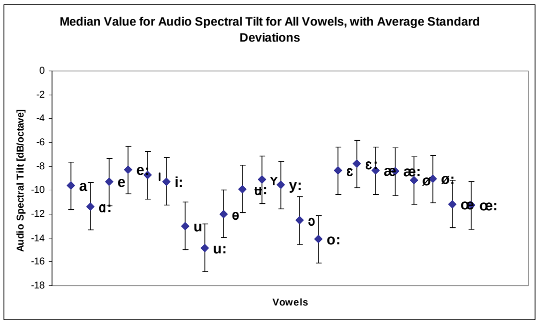

Figure 3. Median values for audio spectral tilt for all vowels, with the average standard deviations.

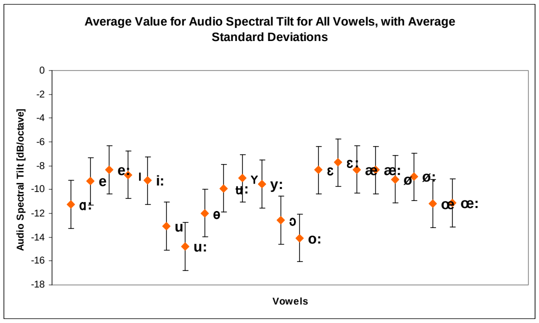

Figure 4. Average values for audio spectral tilt for all vowels, with the average standard deviations.

For EGG spectral tilt, the differences between the vowels are not significant and the small existing variations are shown in the graph with the median values (Figure 5). The average EGG spectral tilt values show a similar pattern (Figure 6).

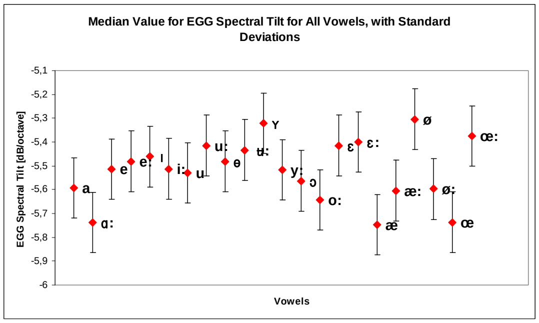

Figure 5. Median values for EGG spectral tilt for all vowels, with the average standard deviations.

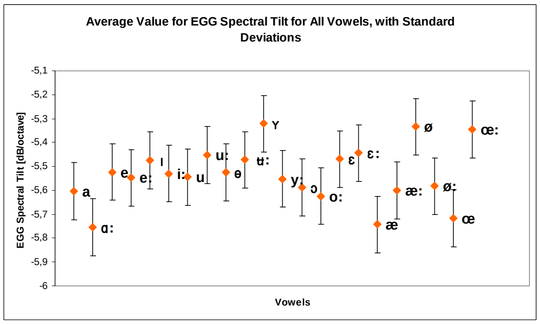

Figure 6. Average values for EGG spectral tilt for all vowels, with the average standard deviations.

The average loudness (Figure 7) and sound pressure level (Figure 8) for each vowel corresponded to each other, with / ʏ / and / œ:/ reaching the highest values.

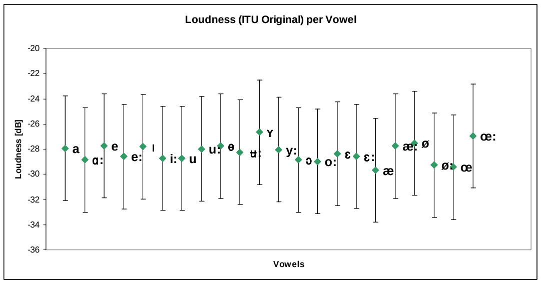

Figure 7. Average values for the ITU Original loudness for all vowels with the average standard deviations.

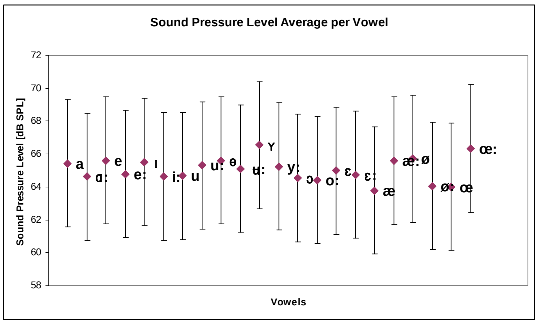

Figure 8. Average values for the sound pressure levels for all vowels with the average standard deviations.

Discussion

This study confirms previous findings of vowel dependence in parameters related to audio spectral tilt and also supports the conclusion that this feature, due to its variability, is insufficient for correct vowel identification. In the light of the source-filter-theory introduced by Fant (1960), it is not surprising that the EGG spectral tilt has no vowel dependence; a vowel is determined by formants which are formed in the vocal tract (the filter) whereas the EGG is related to the source only. This conclusion can, however, be blinded by the fact that we only used one speaker and to problems in EGG measurements, which, although they can be significant, we know nothing about in this case. When using EGG spectral tilt in speech synthesis applications, it is useful to know that vowel dependence is likely to be of little importance. Hence, all sorts of recorded speech, both spontaneous and read, where the vowel distribution is likely to be skewed, can be used without further sorting for EGG spectral tilt.

In this investigation, the sound pressure levels cannot be reconfirmed as being decreased with a steeper audio spectral tilt, also when looking at each vowel separately. This could possibly be explained by a low variability in sound pressure level for the speaker. The EGG spectral tilt, on the other hand, shows a decrease with a lowered sound pressure level.

Further research on the use and nature of the EGG spectral tilt would be valuable, especially since it seems to be a comparatively stable parameter.

Conclusions

The EGG spectral tilt parameter does not vary significantly with vowel type and the values typically get steeper as the sound pressure level increases. The audio spectral tilt values are vowel dependent.

References

Al Moubayed, S, Ananthakrishnan, G & Enflo, L (2010). Automatic Prominence Classification in Swedish. To be published in *Proceedings of Speech Prosody 2010, Workshop on Prosodic Prominence*. Chicago, USA.

Fant, G (1960). Acoustic Theory of Speech Production. Mouton, The Hague, Paris.

Fant, G & Lin, Q (1988). Frequency domain interpretation and derivation of glottal flow parameters. *Speech Transmission Laboratory Quarterly Progress Scientific Report* 1988(2-3): 1-21.

Hanson, H M (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America* 101(1): 466-481.

Hillenbrand, J, Getty, L J, Clark, M J, and Weeler, K (1995). Acoustic characteristics of American English vowels, *J. Acoust. Soc. Am.* 97: 3099-3111.

Iseli, M, Shue, Y-L & Alwan, A (2007). Age, sex and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am.* 121 (4): 2283-2295.

ITU-R (2006). Rec, ITU-R BS. 1770-1, Algorithms to measure audio programme loudness and true-peak audio level. International Telecommunication Union.

Karlsson, I (1988). Glottal waveform parameters for different speaker types. *Speech '88: Proceedings 7th FASE Symposium*, edited by W.A. Ainsworth and J.N. Holmes (Institute of Acoustics, Edinburgh), 225-231.

Karlsson, I (1992). Modelling voice variations in female speech synthesis. *Speech Communication* 11, 1-5.

Kiefte, M & Kluender, K (2005). The relative importance of spectral tilt in monophthongs and diphtongs. *J. Acoust. Soc. Am.* 117: 1395-1404.

Miller, R L (1959). Nature of the vocal cord wave. *J. Acoust. Soc. Am.* 31: 667-677.

Nygren P (2009). Achieving equal loudness between audio files - Evaluation and improvements of loudness algorithms. Master's thesis, Dept. of Speech, Music and Hearing, KTH, Stockholm, Sweden.

O'Haver, T C (1995). Version 2 Last revised Oct 27, 2006: http://terpconnect.umd.edu/~toh/spectrum/findpeaks.m

Orlikoff, R F (1998). The uses and abuses of electro-glottography, *Phonoscope* 1: 37–53.

Rothenberg, M (1973). A new inverse-filtering technique for deriving the glottal airflow during voicing. *J. Acoust. Soc. Am.* 53: 1632-1645.

Sjölander, K (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik 2003*: 93-96.

Timoney J, Lysaght T, Schoenwiesner M, McManus (2004). Implementing loudness models in Matlab. In *Proceedings of the 7th Int. Conference on Digital Audio Effects (DAFX-04)*. Naples, Italy.

Vieira, M N, McInnes, F R & Jack, M A (1996). Robust F0 and jitter estimation in pathological voices. In *ICSLP-1996*: 745-748.