Using phonetic feature extraction to determine optimal speech regions for maximising the effectiveness of glottal source analysis

John Kane, Irena Yanushevskaya, John Dalton, Christer Gobl, Ailbhe Ní Chasaide

Phonetics and Speech Laboratory,
School of Linguistic, Speech and Communication Sciences,
Trinity College, Dublin

kanejo@tcd.ie, yanushei@tcd.ie, jrdalton@tcd.ie, cegobl@tcd.ie, anichsid@tcd.ie

Abstract

Parameterisation of the glottal source has become increasingly useful for speech technology. For many applications it may be desirable to restrict the glottal source feature data to only speech regions where it can be reliably extracted. In this paper we exploit the previously proposed set of binary phonetic feature extractors to help determine optimal regions for glottal source analysis. Besides validation of the phonetic feature extractors, we also quantitatively assess their usefulness for improving voice quality classification and find highly significant reductions in error rates in particular when nasals and fricative regions are excluded.

Index Terms: Glottal source, voice source, phonetic features, voice quality

1. Introduction

The research being carried out by the voice processing group at the Phonetics and Speech Laboratory in Trinity College Dublin is concerned with the development of robust glottal source processing methods and analysing the function of the glottal source in prosody. As part of this endeavour, we have been developing our speech analysis methods to be adaptive and flexible according to the phonetic and prosodic context. This paper illustrates one recent development which involves training a set of feature extractors that derive information relevant to certain binary phonetic classes. This information is then exploited in order to improve the effectiveness of glottal source analysis by allowing sensitivity to the phonetic context.

Parametric characterisation of the glottal source component of speech offers a rich source of information often not captured with standard spectral features. However, glottal inverse filtering (the process of estimating the glottal source by compensating for the effect of vocal tract resonance) typically displays sub-optimal performance in speech involving certain phonetic classes. For instance, nasal consonants and nasalisation in general involves zeros in the vocal tract spectrum which are not characterised by the commonly used all-pole vocal tract model. Subsequent parameterisation of certain aspects of the glottal source (in particular the return phase) has been shown to be significantly affected by this [1]. The interaction of a low first formant frequency (F1; commonly found in high vowels) with a high fundamental frequency (f_0) is challenging for automatic glottal inverse filtering, particularly in terms of discriminating F1 from the glottal formant [2, 3]. Voiced fricatives are also extremely challenging for both inverse filtering and glottal source parameterisation procedures.

For certain applications, for instance fully parametric speech synthesis [4, 5], complete modelling of the glottal source is required for all (voiced) regions of the speech data. For many other applications, such as speaker identification [6], voice quality detection [7, 8] and emotion classification [9, 10, 11], parameterisation may not be required for every region in the speech signal and these applications may indeed benefit from slightly less glottal feature data which has a higher likelihood of being reliable. A previous study [12] proposed a method of determining *centres of reliability*, which were defined as vocoids with strong sonorant energy, in a spectrally steady region where reliable formant extraction is possible. Although this approach has the potential for improving the effectiveness of glottal source analysis it was not formally evaluated for this purpose.

In the present study, we propose the design of a set of independent feature extractors for binary phonetic classes, and using this information to determine optimal regions for glottal source analysis. Such an approach has been proposed previously [13], and has promising implications for both speech synthesis [14] and recognition [15]. However, it has not yet been exploited in terms of glottal source analysis. Training and validation of such phonetic feature extractors is described in the current paper. Their effectiveness in optimising glottal source analysis is objectively assessed through the use of voice quality classification experiments, where the effect of excluding certain speech regions, as well as combining these phonetic features within the parameter set, on classification accuracy levels is quantified.

2. Phonetic feature extraction

2.1. Speech data

In order to train and validate the approach of extracting phonetic features, a reasonably large set of data is required including a wide phonetic coverage as well as involving a range of speakers. We opt to use freely available databases containing phonetic transcriptions (see Table 1). The first is the ARCTIC database [16] which contains 9 speakers of various English dialects and the second is the newly available IIIT Indic speech database [17], from which we use data from 6 speakers.

In the present study we investigate four binary phonetic feature classes: voicing, frication, nasals and high vowels (chosen due to their relevance to glottal source analysis). Using the labels available with these databases we create binary vectors for each feature which are subsequently used as targets in the detection training. Say for instance for frication, all the regions labelled as being fricatives (e.g., /f/, /v/, /s/, etc.) are allocated the target 1, with all other labels assigned 0. Note that in the present data only nasal consonants were present, no nasalised

Table 1: Summary of speech data used in training and validating phonetic feature extraction.

Data	Speaker ID	Language	Gender	Utterances
ARCTIC	AWB	English	Male	1138
	BDL	English	Male	1142
	CLB	English	Female	1132
	JMK	English	Male	1132
	KED	English	Male	452
	KSP	English	Male	1132
	RAB	English	Male	1946
	RMS	English	Male	1132
	SLT	English	Female	1132
IIIT	ABI	Malayalam	Male	1000
	ANT	Bengali	Male	1000
	ASH	Marathi	Male	1000
	LP	Kannada	Female	1000
	MOH	Tamil	Male	1000
	SUKH	Hindi	Female	1000

vowels. For future work, we intend on developing a phonetic feature extractor for nasality in general, with the method described in the current work being limited to nasal consonants.

2.2. Acoustic features and Artificial Neural Networks

The acoustic features used in the present study are the standard Mel-frequency cepstral coefficients (MFCCs) measured on 25 ms Hanning windowed frames with a 10 ms shift. Note that the first MFCC, related to signal energy, is normalised to the maximum energy for a given utterance. Δ and $\Delta\Delta$ coefficients are also included, resulting in a 39-dimensional feature vector.

In order to detect the phonetic features we use this feature vector as input to artificial neural networks (ANNs). In general ANNs are used to learn a mapping f from the input feature space I (typically $\mathbb{R}^n$, i.e. the above described features) to the target space T (in this case $\{0, 1\}$, i.e. the individual phonetic feature binary target): $f(\mathbf{x}) : \mathbf{x} \in I \rightarrow \mathbf{y} \in T$, where $\mathbf{x}$ denotes the input vector and $\mathbf{y}$ the output of the approximator f.

We chose the well known multi-layer perceptron (MLP) as the network type of choice [18] which is computationally inexpensive once the network parameters are trained. The network architecture is designed as a two layer MLP with one hidden layer containing 100 neurons, all fully connected to the input and the output layer. The hidden layer neurons use $\tanh$ as a transfer-function and the output neuron uses a linear transfer-function, as this is optimal for function approximation. The training is conducted using a standard error back-propagation algorithm [19]. As the binary phonetic features: high vowels, frication and nasals, have a rather sparse occurrence, the decision threshold, θ, is set for these individual feature extractors by maximising the F1 score (see below) on the training set. For the voicing feature extractor, which does not have a sparse occurrence, θ is set at 0.5.

In order to validate this approach for extracting phonetic features, we use the data summarised in Table 1. Leave one speaker out validation is used, whereby data from one speaker is held out to be used solely for testing while the rest of the data is used for training the ANNs. The process is repeated for each speaker. Classification error (i.e. the percentage of speech frames misclassified) is used as an evaluation metric. However, for sparsely occurring events (e.g., nasals, which may only occur in 5 % of the speech data) classification error is unsuitable. To address this we also include the F1 score, which takes into account true positives (Tp), false positives (Fp) and false negatives (Fn) and is a more robust metric for skewed datasets:

$$F1 = \frac{2 \cdot \mathrm{Tp}}{2 \cdot \mathrm{Tp} + \mathrm{Fp} + \mathrm{Fn}} \in [0, 1] \quad (1)$$

Roughly speaking, above 0.7 indicates high accuracy, around 0.5 indicates moderate accuracy and below 0.3 indicates low accuracy.

Table 2: Leave one speaker out validation results for phonetic feature extraction training and testing. Detection error and F1 score are given summarised as means ($\hat{x}$) and standard deviation (σ).

Phonetic feature	Error (%)		F1	
	$\hat{x}$	σ	$\hat{x}$	σ
Voicing	16.8	3.1	0.89	0.02
Frication	11.1	2.7	0.46	0.16
Nasals	10.6	2.14	0.42	0.13
High vowels	15.7	4.6	0.47	0.11

2.3. Results

The results from validation of the phonetic feature extraction method are displayed in Table 2. Unsurprisingly, the detection accuracy of voicing is very high with an F1 score of 0.89. Frication and nasals are both detected with reasonably low detection error rates (approx. 11 %), however with a somewhat lower F1 score. For frication, the vast majority of false positives are voiceless stops (/t/, /k/). In effect many of these are probably not true false positives, as the extractor is detecting the aspiration often accompanying these stops as well as allophonic lenition which entails these sounds being produced as fricatives. Also, for nasals a large proportion of the false positives are laterals which display a high degree of acoustic similarity to nasals. High vowels are detected with a slightly higher error rate, but with also a slightly higher F1 score (0.47).

Although there is clearly potential for improvement in the accuracy of these extractors, which may be brought about by better parametric descriptions of the speech specifically relevant to these phonetic classes, they nevertheless, in their current form, provide valuable phonetic information. This information may be beneficial for, among other applications, optimising glottal source analysis. An illustration showing the output of these feature extractors is shown in Figure 1. The contours above the spectrogram give information on the presence of the phonetic classes: voicing, frication, nasals and high vowels. For instance, consider the contour for nasals in the two nasal regions at 0.1 and 1.8 seconds, which suitably show high confidence of nasals. Similarly, the fricatives at around 0.55 and 1.5 seconds show up with high values indicating high confidence of the presence of frication.

3. Voice quality classification

Explicit objective evaluation of the effectiveness of glottal source analysis is far from straightforward, particularly as it is

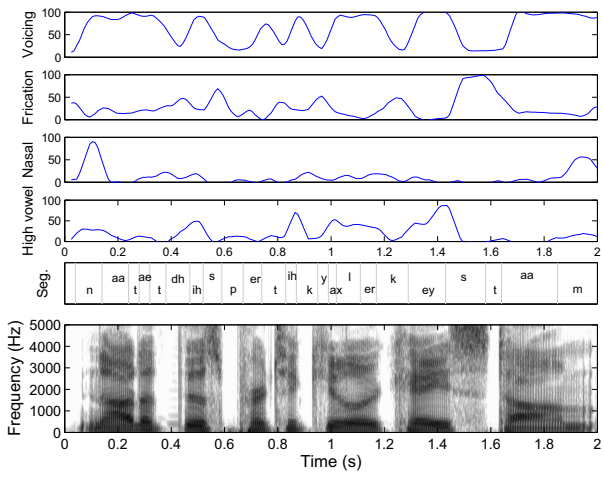

Figure 1: Broadband spectrogram of the sentence *Not at this particular case Tom, ...* (bottom panel), said by an American male. The top three panels show confidence of the presence of voicing, frication, nasals and high vowels. The fourth panel shows the ARCTIC transcription.

not possible to obtain strictly reliable reference values from natural speech. In order to objectively evaluate the glottal source analysis in the present study we propose voice quality classification experiments with the assumption being that the effectiveness of the analysis will be implicitly shown through the accuracy of the voice quality classification.

3.1. Speech data

In order to evaluate the ability of the phonetic feature extractors to improve the effectiveness of glottal source analysis we required a reasonably large database of speech produced in different voice qualities (note that such databases are not widely available). For this, we opt to use a section of the speech data used in [20]. This consists of speech produced by 6 speakers (3 male and 3 female, all experienced in speech research) recorded in a semi-anechoic chamber. Audio was captured using high quality recording equipment: a B & K 4191 free-field microphone and a B & K 7749 pre-amplifier. The signals were digitised at 44.1 kHz (using a LYNX-two sound card) and were subsequently downsampled to 16 kHz. Participants were asked to read 17 sentences in six different phonation types (though only the breathy, modal and tense samples were used here). The sentences were chosen from the *phonetically compact* sentences in the TIMIT corpus, four of which contained all-voiced sounds. These sentences were chosen in order to obtain a reasonably wide phonetic coverage. Participants were given prototype voice quality examples, produced by John Laver [1] and the present first author, and were asked to practise producing them before coming to the recording session. For the recordings participants were asked to produce the versions of each phonation type with emphasis and to maintain it throughout the utterance. During the recording session participants were asked to repeat the sentence when it was deemed necessary. Note that none of the speech data used in the development of the phonetic feature extractors was used in the voice quality classification experiments.

[1] These examples come as part of Laver (1980)

3.2. Glottal source parameters

We use glottal source parameters derived both from direct measures of the estimated glottal source signal as well as parameters derived following the fitting of a model to the glottal source data. Glottal closure instants (GCIs) are first detected using the SE-VQ algorithm [20]. Glottal inverse filtering is then carried out using the iterative and adaptive inverse filtering (IAIF; [21]) method. Parameterisation of the glottal source estimate is then carried out by fitting the Liljencrants-Fant (LF) glottal source model [22] using the recently proposed DyProg-LF method which has been shown to be more robust compared to traditional methods [23]. The DyProg-LF method, employs a dynamic programming algorithm to determine an optimal path of Rd values, the global shape parameter of the LF model, through the speech signal. There then follows an optimisation procedure to refine the model fit. The full set of parameters derived from this procedure includes EE, which is the strength of the main excitation, Ra, which characterises the return phase, Rk, which describes the asymmetry of the glottal pulse, and Rg which relates to the normalised frequency of the glottal formant. For the present study we exclude EE (as we would like to detect voice quality variation independent of energy related measurements) and include Rd giving the set: $\{f_0$, Ra, Rk, Rg, Rd$\}$.

Parameterisation is also carried out through the use of direct measures. We selected three parameters which were previously shown to be effective for discriminating breathy, modal and tense voice [24]. The normalised amplitude quotient (NAQ, [25]) is calculated with:

$$\text{NAQ} = \frac{f_{ac}}{d_{peak} \cdot T_0} \tag{2}$$

where f_{ac} is the maximum amplitude of the glottal flow pulse and d_{peak} is the amplitude of the maximum negative amplitude of the glottal derivative pulse (see Figure 2). The quasi-open quotient (QOQ, [26]) was developed as more robust measure relating to the standard open quotient. It is calculated by detecting the peak in the glottal flow and finding the time points previous to and following this point that descend below 50 % of the peak amplitude (see Figure 2). The duration between these time locations is used as a 'quasi-open phase' and is divided by the local glottal period in order to derive QOQ. The difference between the first two harmonics (H1-H2) in the narrowband amplitude spectrum of the glottal source signal is also included. This is calculated on GCI centred glottal source frames, three times the local glottal period in duration. Combining both direct and model based parameters provides a 8-dimensional feature vector: $\{f_0$, Ra, Rk, Rg, Rd, NAQ, QOQ, H1-H2$\}$.

3.3. Classification experiments

The above described glottal source parameters, derived from the speech data described in Section 3.1, are used as input features to a Support Vector Machines (SVM) classifier, using a radial basis function kernel and a one-against-one multi-class architecture. The training targets are the voice quality labels: breathy, modal and tense voice. 10-fold cross-validation experiments are conducted where the voice quality dataset is randomly partitioned into 10 equal-sized sets. One of the sets is held out for testing, with the remainder used for training, and this is repeated for each of the 10 sets. Note that due to the rather low number of speakers (due to the lack of widely available voice quality labelled data) we decided not to include speaker independent validation experiments. Phonetic feature extraction described in Section 2 (and trained on all the data displayed

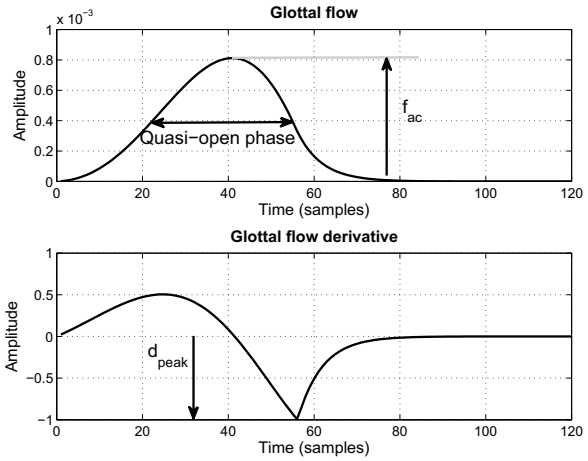

Figure 2: Glottal flow (top) and glottal flow derivative pulse (bottom) with the measurements required for deriving NAQ (f_{ac} and d_{peak}) and QOQ (quasi-open phase) highlighted.

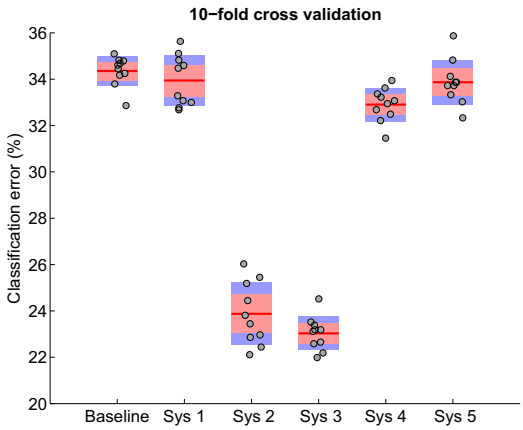

Figure 3: Distribution of voice quality classification errors in the 10-fold cross validation experiments, plotted as a function of system type. The plot shows individual data points, the mean (red line), 95 % confidence interval (red shading) and standard deviation (blue shading) of each distribution.

in Table 1), is carried out on the voice quality speech data and this information is utilised in the classification experiments.

The classification experiments involve assessment of five different systems, which are designed to test the usefulness of the phonetic feature extractors for optimising the effectiveness of the glottal source analysis:

Baseline: Using all the glottal source parameters derived from voiced speech regions.

System 1: Baseline system, excluding glottal source parameters detected in high vowel regions.

System 2: System 1, additionally excluding glottal source parameters detected in fricative regions.

System 3: System 2, additionally excluding glottal source parameters detected in nasal regions.

System 4: Baseline system, only excluding glottal source parameters detected in nasal regions.

System 5: Baseline system, incorporating the four phonetic features derived using ANNs as additional input features.

3.4. Results

The results of the 10-fold cross validation experiments are plotted as a function of system type in Figure 3. A one-way ANOVA, with classification error treated as the dependent variable and system type as the independent variable, reveals a highly significant effect of system type [$F_{(5,54)} = 317.2$, $p < 0.001$]. In terms of the median classification errors we observe the following trend:

baseline $>$ system 1 $\gg^*$ system 2 $>$ system 3,

where * indicates a significant difference (at $p < 0.001$, following Tukey's Honestly Significant Difference (HSD) test). We can interpret from this that excluding high vowel regions, as detected by the phonetic feature extractor, only brings a minor improvement to the voice quality classification. Excluding areas of frication brings a dramatic reduction in the classification error and further exclusion of nasals brings a further minor improvement in voice quality classification. Considering just exclusion of detected nasal regions (i.e. system 4) we observe a significant ($p < 0.05$) reduction in classification error over the baseline. The effect of removing areas detected as containing frication and nasals clearly result in an improvement in the accuracy of voice quality classification.

For system 5, which includes the extracted phonetic feature contours as extra features in the classifier training offers a slight, but not significant drop in error compared to the baseline system.

4. Discussion & conclusion

This paper introduces a novel approach to determining optimal speech regions for glottal source analysis, by exploiting information provided by a set of binary phonetic feature extractors. The effectiveness of glottal source analysis is assessed via a voice quality classification experiment. This experiment reveals a dramatic and highly significant reduction in classification error, particularly when regions detected as containing frication or nasals are excluded. Although we have not assessed the effectiveness of glottal source parameterisation directly, these findings are indicative of more reliable glottal source modelling. This development may have important implications for a range of speech processing applications. For future work, we intend to extend the set of extractors to cover a richer set of phonetic features. We are in the process of developing acoustic features for specific phonetic classes, e.g., nasals, and we hope to utilise these, along with standard spectral features, to improve the generalisablilty of these phonetic feature extractors. Finally, we intend to utilise such phonetic information to optimise and create more adaptive glottal inverse filtering.

5. Acknowledgements

This research is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET) and the Irish Department of Arts, Heritage and the Gaeltacht (ABAIR project). The authors would also like to thank Stefan Scherer for his help with the artificial neural networks and support vector machines classifier.

6. References

[1] Gobl, C., Mahshie, J., (2013) "Inverse filtering of nasalized vowels using synthesized speech", Journal of Voice, 27(2), 155-169.

[2] Gobl, C., (1988) "Voice source dynamics in connected speech", KTH, Speech Transmission Laboratory, Quarterly Report, 29, 123-159.

[3] Walker, J., and Murphy, P., (2007) "A review of glottal waveform analysis" *in* Progress in nonlinear speech processing, pp. 1-21.

[4] Cabral, J., Renals, S., Richmond, K., Yamagishi, J., (2011) "HMM-based speech synthesiser using the LF-model of the glottal source", Proceedings of ICASSP, Prague, Czech Republic, 4704-4707.

[5] Degottex, G., Lanchantin, A., Roebel, A., Rodet, X., (2012) "Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis", Speech Communication, 55(2), pp. 278-294.

[6] Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S., Jameel, H., Richey, C., Goodman, F., (2008) "Effects of vocal effort and speaking style on text-independent speaker verification", Proceedings of Interspeech, Brisbane, Australia, 609-612.

[7] Scherer, S., Kane, J., Gobl, C., Schwenker, F., (2013) "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification", Computer Speech and Language 27(1), 263-287.

[8] Kane, J., Gobl, (2013) "Wavelet maxima dispersion for breathy to tense voice discrimination", IEEE Transactions on Audio, Speech and Language Processing, 21(6), 1170-1179.

[9] Lugger, M., (2007) "The relevance of voice quality features in speaker independent emotion recognition", Proceedings of ICASSP, Hawaii, USA, 17-20.

[10] Yanushevskaya, I., Gobl, C., Ní Chasaide, A., (2009) "Voice parameter dynamics in portrayed emotions", Proceedings of Maveba, pp. 21-24.

[11] Lliev, A. I., Scordilis, M., Papa, J., Falcão, A., (2010) "Spoken emotion recognition through optimum-path forest classification using glottal features", Computer Speech and Language 24(3), pp. 445-460.

[12] Mokhtari, P., Campbell, N., (2002) "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech", Proceedings of LREC, 2015-2018.

[13] Amer, T-A, Carson-Berndsen, J., (2003) "Hartfex: A multi-dimensional system of HMM based recognizers for articulatory feature extraction," in Proceedings of the 8th EUROSPEECH, Geneva, Switzerland.

[14] Cahill, P., Aioanei, D., Carson-Berndsen, J., (2007) "Articulatory acoustic feature applications in speech synthesis" Proceedings of Interspeech, Antwerp, Belgium, 2877-2880.

[15] Kane, M., Mauclair, J., Carson-Berndsen, J., (2011) "Automatic identification of phonetic similarity based on underspecification" Human Language Technology. Lecture notes in Computer Science, 6563, 47-58.

[16] Kominek, J., Black, A., (2004) "The CMU ARCTIC speech synthesis databases" ISCA speech synthesis workshop, Pittsburgh, PA, 223-224.

[17] Prahallad, K., Kumar, E., Keri, V., Rajendran, S., Black, A., (2012) "The IIIT-H Indic speech databases" Proceedings of Interspeech, Portland, Oregon, USA.

[18] Hornik, K., (1991) "Approximation capabilities of multilayer feedforward networks", Neural Networks, Elsevier, 4, 251-257.

[19] Bishop, C. M., (2006) Pattern Recognition and Machine Learning (Information Science and Statistics) Springer.

[20] Kane, J., Gobl, C., (2013) "Evaluation of glottal closure instant detection in a range of voice qualities", Speech Communication 55(2), 295-314.

[21] Alku, P., (1992) "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Communication, 11, 109-118.

[22] Fant, G., Liljencrants, J., Lin, Q., (1985) "A four parameter model of glottal flow" KTH, Speech Transmission Laboratory, Quarterly Report, 4, 1-13.

[23] Kane, J., Gobl, C., (2013) "Automating manual user strategies for precise voice source analysis", Speech Communication 55(3), 397-414.

[24] Airas, M., Alku, P., (2007) "Comparison of multiple voice source parameters in different phonation types", Proceedings of Interspeech Antwerp, Belgium, 2007, 1410-1413.

[25] Alku, P., Bäckström, T., Vilkman, E., (2002) "Normalized amplitude quotient for parameterization of the glottal flow", Journal of the Acoustical Society of America, 112, 701-710.

[26] Hacki, T., (1989) "Klassifizierung von Glottisdysfunktionen mit Hilfe der Elektroglottographie", Folia phoniatrica, 41, 43-48.