Glottal inverse filtering and its application in automatic classification of diseases

Paavo Alku

Aalto University, Finland

Department of Information and Communications Engineering

Contents

1. Introduction

Basics of glottal inverse filtering (GIF)

(a) Production of (voiced) speech is modelled by three processes:

(1) Glottal flow: excitation

(2) Vocal tract: prominent physiological filter with resonances (i.e. formants)

(3) Lip radiation: conversion of flow in mouth into pressure in free field

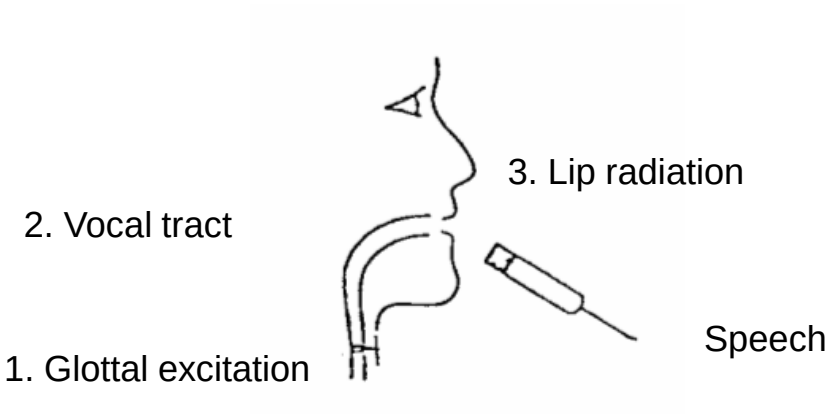


Fig. 1. Production of speech.

(b) Filtering processes (vocal tract, lip radiation) are estimated and represented typically by digital filters. Lip radiation is modelled by a fixed high-pass filter, whereas modelling of the vocal tract is more challenging and calls for adaptive methods.

(c) Glottal excitation is estimated by cancelling the effects of the vocal tract and lip radiation by filtering speech through the inverses of the filters.

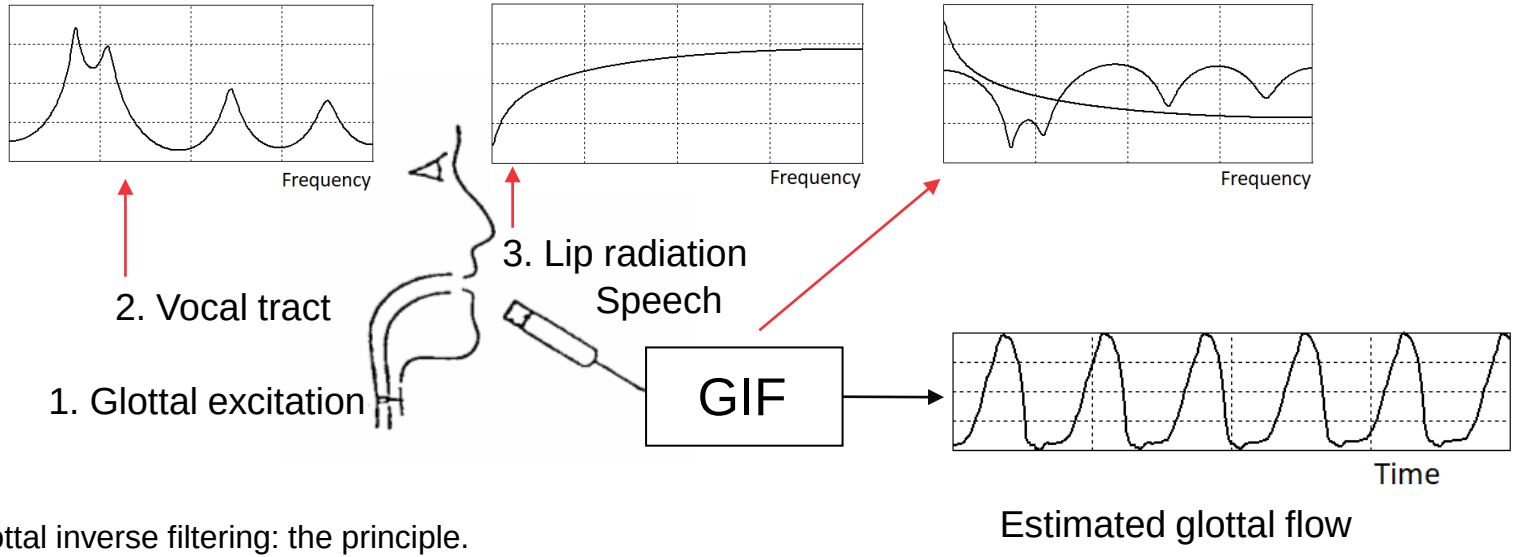


Fig. 2. Glottal inverse filtering: the principle.

- In the current talk, we will focus on the inverse filtering scenario shown in Fig. 2 that is based on using as input the speech pressure signal recorded outside the mouth.
- Inverse filtering can also be computed using the oral flow input recoded with a mask (Rothenberg, 1973) and this approach has been used in many studies (e.g. Sundberg, 2017; Sundberg *et al*., 2005).
- The use of free-field input enables applying inverse filtering in biomarking state of health from speech recordings conducted in real-life situations (e.g. from spontaneous conversations).

Benefits of inverse filtering

- Provides a **non-invasive** method to estimate the origin of voiced speech.
- Reveals **valuable information** about the generation of several essential acoustical cues (e.g. F0, phonation type) that are used in speech communication. Differently from other techniques such as electroglottography this information corresponds to **real acoustical** phenomenon (air flow) that takes place when speech is produced.
- Can be used in **technical applications** such as in automatic classification of disorders and in speech synthesis.

Drawbacks of inverse filtering

- Estimation **accuracy** sometimes not satisfactory (e.g., for high-pitched speech).
- Some GIF methods need **user adjustments** => Estimation results might be biased by the user.
- Calls for high-quality **recording equipment.**
- Many GIF methods are based on linear source-tract models => Non-linear phenomena in speech production (e.g. interaction between source and tract) cannot be modelled.

Voice source analysis is typically conducted in two stages

(1) The inverse filtering stage

- Input: pressure signal recorded by microphone
- Output: time-domain waveform of the estimated glottal excitation

(2) The parameterization stage

- Input: time-domain waveform of the estimated glottal excitation
- Output: numerical value(s) that captures the most essential information embedded in the glottal excitation waveform

2. Glottal inverse filtering methods

- The idea of GIF was proposed in Miller (1959).

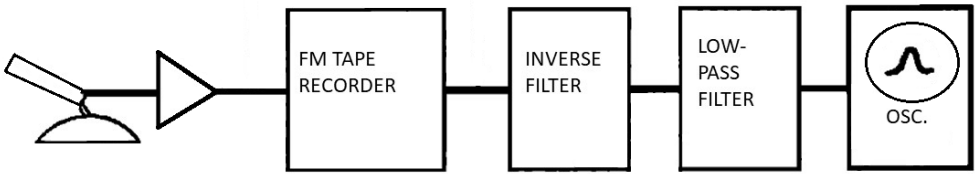


Fig. 3. Schematic of the experimental GIF setup used in Miller (1959).

- Other early experiments using analog techniques (e.g. Fant 1961; Lindqvist-Gauffin 1965; Rothenberg, 1973)
- First GIF experiments based on digital signal processing were conducted by Oppenheim and Schafer (1968) and Nakatsui and Suzuki (1970).
- Many GIF methods have been proposed in the past four decades, examples:
  - Closed phase analysis (Strube, 1974; Wong *et al*., 1979)
  - Iterative adaptive inverse filtering (IAIF) (Alku, 1992)
  - Simultaneous inverse filtering and model matching (Fröhlich *et al*., 2001)
  - Zeros of z-transform (Bozkurt *et al*., 2005)
  - Autoregressive model with an exogenous input (Fu and Murphy, 2006)
  - Complex cepstrum based decomposition (Drugman *et al*., 2011)
  - **Quasi-closed phase analysis** (Airaksinen *et al*., 2014)
  - State-space modelling optimized by Kalman filtering (Sahoo and Routray, 2016; Cortes *et al*., 2022) or the expectation-maximization algorithm (Alzamendi and Schlotthauer, 2017)
  - Quadratic programming (Airaksinen, Bäckström and Alku, 2017)
  - Modified IAIF (Mokhtari *et al*., 2018)
  - Deep neural network –based analysis (Narendra *et al*., 2019; Langheinrich *et al*., 2022)

3. Glottal parameters

- Glottal inverse filtering maps a time-domain speech waveform to a time-domain glottal flow estimate.

  =>

  The estimated time-domain waveform is expressed using glottal parameters (*i.e*. numerical values) that aim to capture the most essential information embedded in the glottal excitation waveform.

- Glottal parameters are used both in fundamental research of voice production and also in machine learning –based classification studies (as will be described in Section 5).

- Many glottal parameters have been developed both using time-domain and frequency-domain approaches.

  Examples of **time-domain** glottal parameters: (1) based on extracting critical time instants: open quotient, speed quotient (Timcke *et al*., 1958), closing quotient (Monsen and Engebretson, 1977), normalized amplitude quotient (Alku *et al*., 2002), (2) based on matching the estimated flow (or its derivative) with artificial waveforms (e.g. the Liljencrants-Fant model) (Strik and Boves, 1992).

  Examples of **frequency-domain** glottal parameters: H1-H2 (Titze and Sundberg, 1992); harmonic richness factor (Childers and Lee, 1991), parabolic spectral parameter (Alku *et al*., 1997)

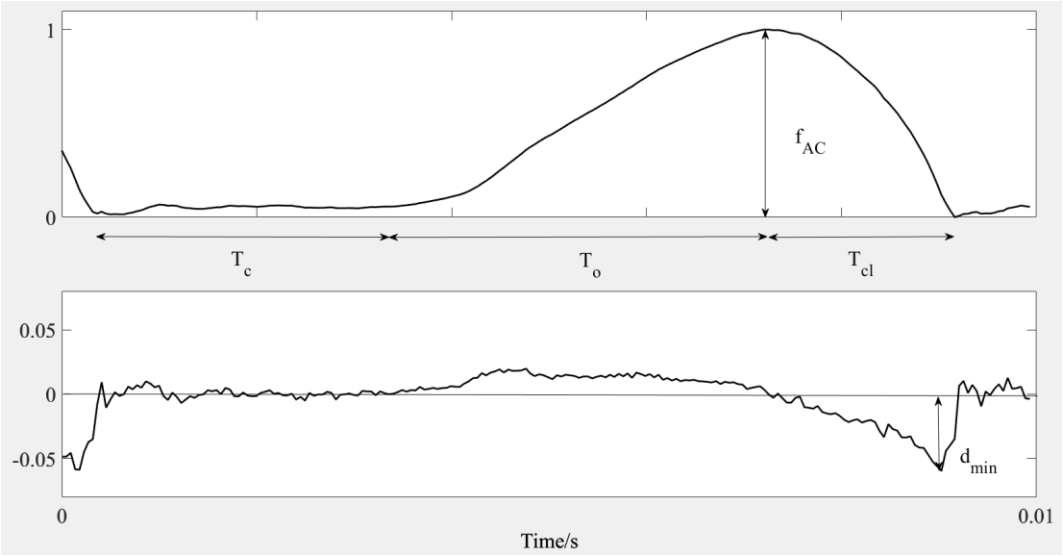


Fig. 4. Glottal flow pulse (upper panel) and its derivative (lower panel).

An example of a time-domain glottal parameter (closing quotient): $ClQ = T_{cl} / (T_c + T_o + T_{cl})$

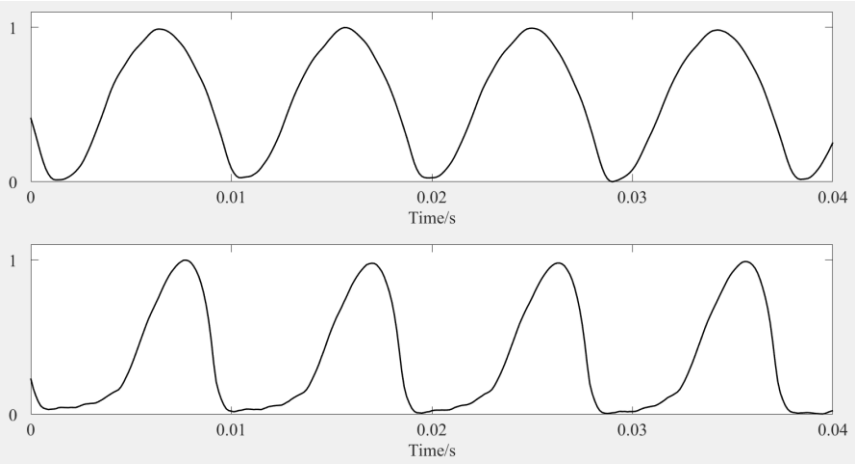


Fig. 5. Time-domain glottal flow pulseform in breathy phonation (upper panel) and in pressed phonation (lower panel).

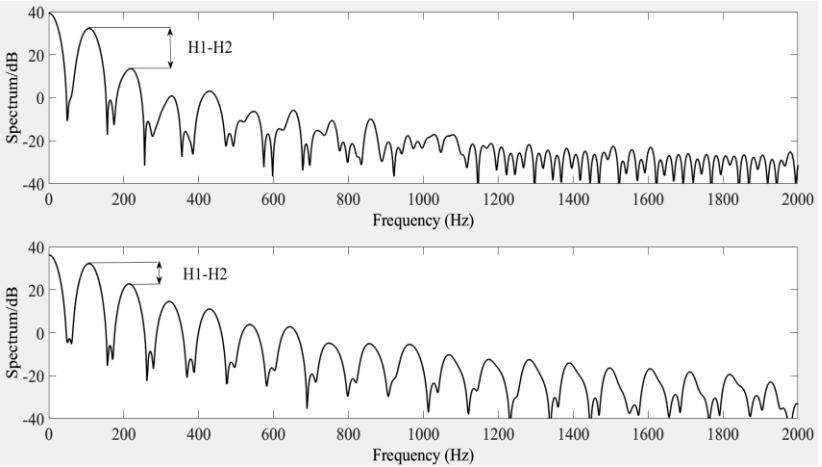


Fig. 6. Glottal flow spectrum in breathy phonation (upper panel) and in pressed phonation (lower panel).

An example of a frequency-domain glottal parameter (level difference at F0 and 2F0): H1-H2

4. Automatic machine learning -based biomarking of health from speech

- Speech is produced by the human speech production mechanism and therefore it contain acoustical cues of the speaker's state of health.
- Signal processing and machine learning (ML) can be used to automatically biomark state of health from speech. In this talk, the technology is called "Speech-based biomarking of health":

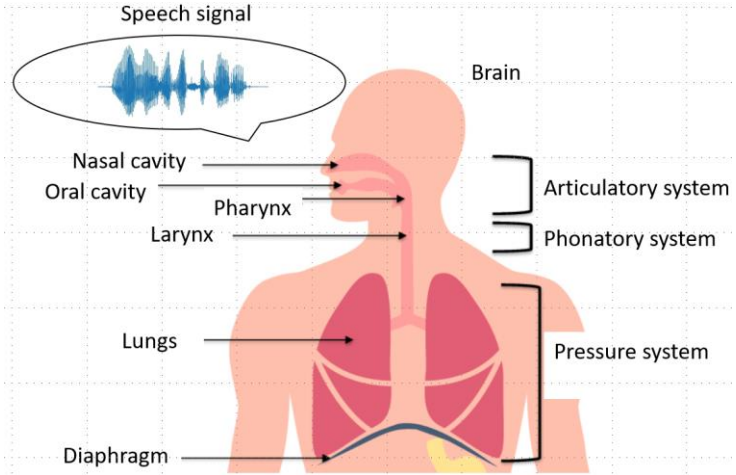


Fig. 7. Organs affecting acoustical cues embedded in speech

Typical machine learning tasks studied

- The **binary classification problem**: detection of speakers with, for example, spasmodic dysphonia from healthy controls
- The **multiclass classification problem**: classification of speakers with different diseases (*e.g*. healthy *vs*. hyperfunctional dysphonia vs. vocal fold paresis) or classification of speakers with different levels of disease severity (e.g. healthy vs. mild dysarthria vs. severe dysarthria)

This presentation focuses on the binary classification problem. (1) Two technologies (the classical pipeline approach, the end-to-end approach) are first generally described in this section. (2) Section 5 presents two examples of studies where automatic ML-based detection of diseases has been investigated using inverse filtering.

4.1 The classical pipeline approach

- Two separate parts: feature extraction and classifier
- Speech microphone signals are expressed in a compressed form using **features** such as mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980).
- A machine learning (ML) **classifier** (e.g. support vector machine, SVM, Cortes and Vapnik, 1995) is trained to distinguish, for example, speech produced in disease “X” from speech produced by healthy speakers.

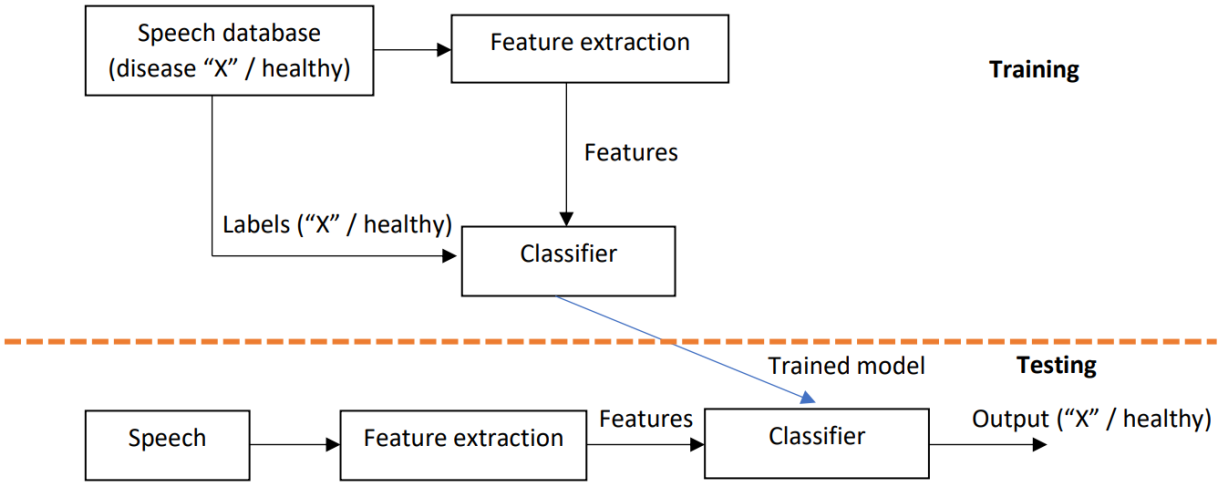


Fig. 8. A general structure of a detection system based on the traditional pipeline approach. Speech database includes labelled speech signals (disease “X” *vs*. healthy). Classifier can be, for example, SVM.

The support vector machine (**SVM**) classifier

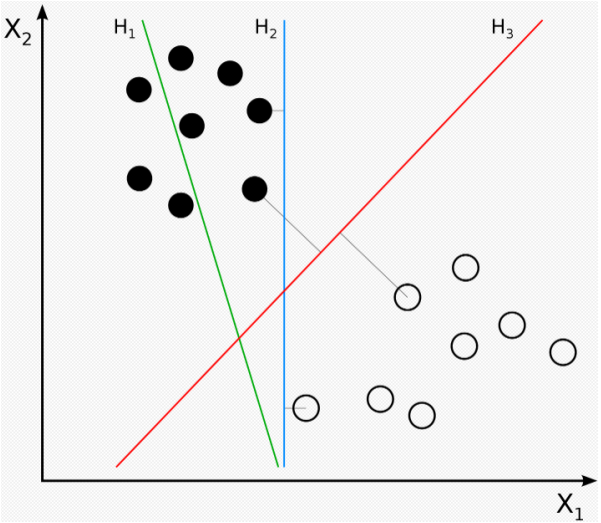


Fig. 9. **Training** data in two classes (back circles: disease “X”; white circles: healthy) in a 2-dimensional speech feature space. SVM defines H_3, which separates the classes with the maximal margin.

- **Test samples** are mapped into that same space and predicted to belong to a category based on which side of the gap they fall.
- Because the two classes in the original, low-dimensional feature space are not necessarily linearly separable, SVM implicitly maps the data to a higher-dimensional space using a **kernel** function.

4.2 The end-to-end approach

- Detection is computed directly from the time-domain input speech waveform (or from spectrogram) using deep learning networks (typically convolutional neural nets, CNNs).
- No separate feature extraction and classification stages: they are replaced with a single network structure.

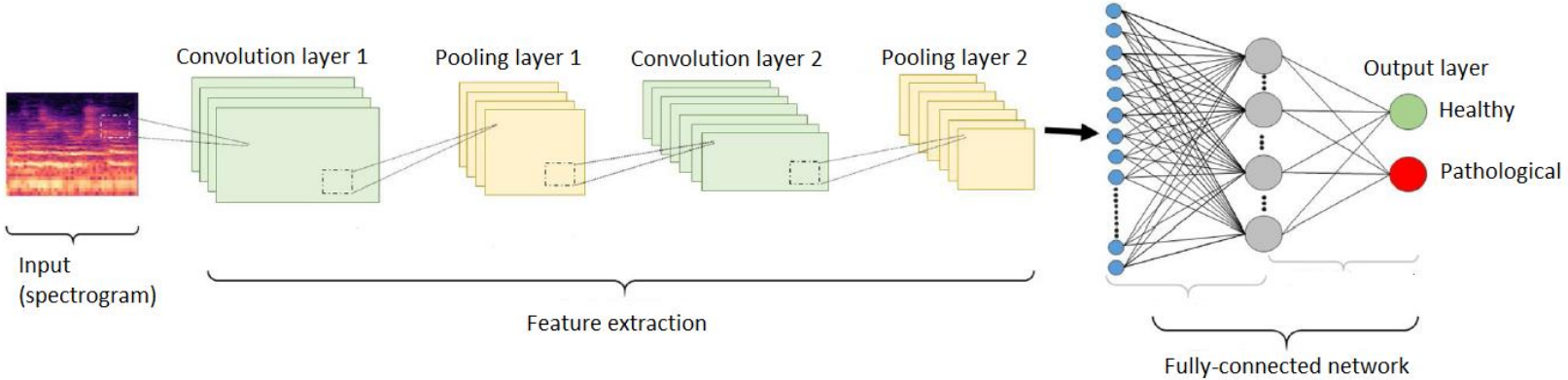


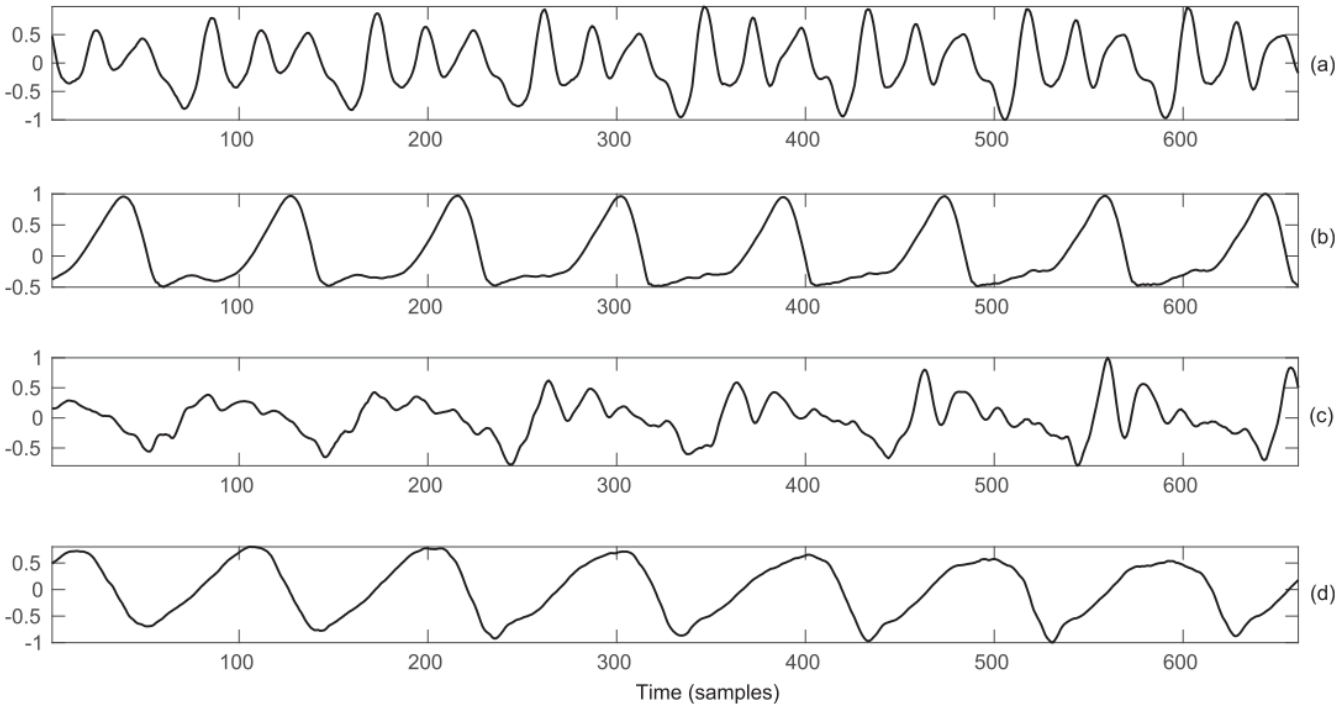
Fig. 10. An example of a CNN-based deep learning classification system (binary detection between “healthy” and “pathological” voice) (modified from Kaushik *et al.*, 2021).

5. Machine learning studies in automatic classification of diseases using glottal inverse filtering

- Automatic ML-based classification of diseases from speech has been widely studied in recent years.
- Most studies use existing feature sets, such as openSMILE (Eyben *et al*., 2010), that are extracted from speech pressure signals.
- Glottal features have been used in some studies, for example:
  - Classification of nodules in children (Szklanny and Wrzeciono, 2019)
  - Pathology classification (Wu *et al*., 2021; Gomez-Vilda *et al*., 2009; Kadiri and Alku, 2020; Tirronen *et al*., 2023)
  - Detection of Parkinson's disease (Novotny *et al*., 2020; Vasquez-Correa *et al*., 2021; Narendra *et al*., 2021; Liu *et al*., 2023;
  - Detection of Covid-19 (Deshmukh *et al*., 2021)
  - Detection of depression (Ooi *et al*., 2012; Simantiraki *et al*., 2017)
- Let us study next two of our investigations where glottal features have been used in automatic ML-based detection of diseases from speech.

5.1 Reddy, Helkkula, Keerthana, Kaitue, Minkkinen, Tolppanen, Nieminen, Alku. The automatic detection of heart failure using speech signals. Comp. Speech Lang. 69, article no. 101205, 2021.

- Binary detection: heart failure (HF) vs. healthy
- Speech data (Finnish): 20 patients, 25 healthy controls (recorded by the team at Helsinki University Central Hospital)
- System architecture: a classical pipeline system (with different classifiers and features)
- Study goal: can the detection of HF be improved by using **glottal** features?



Glottal flow (healthy speaker)

Glottal flow (HF patient)

Fig. 11. Speech signal (a) and the estimated glottal flow (b) of a healthy speaker. Speech signal (c) and the estimated glottal flow (d) of an HF patient.

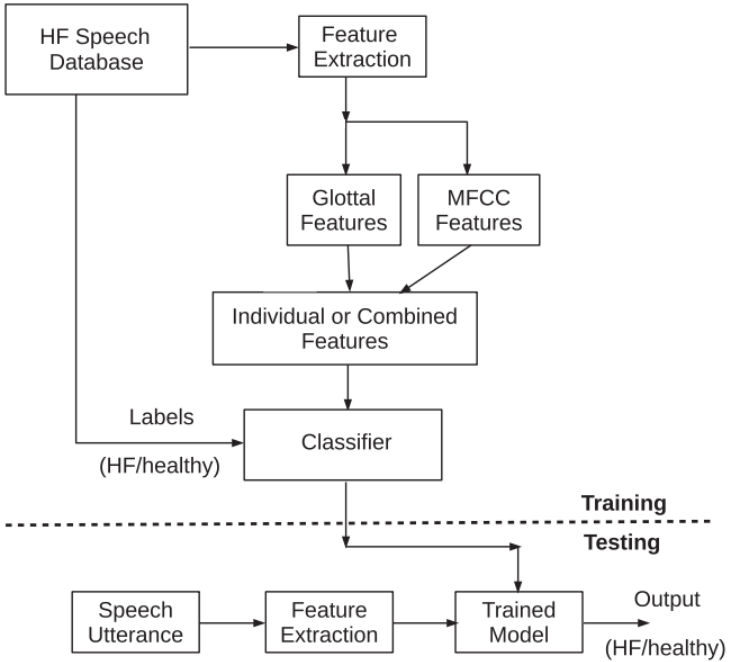


Fig. 12. The classical pipeline method utilized to detect HF from speech.

Baseline: MFCC features and SVM classifier

The best system: combined glottal and MFCC features and neural net classifier

Classifier	MFS (104)	GFS (192)	MFS + GFS (296)	MFS + GFS with feature selection (85)
SVM	76.26	65.59	75.52	77.15
ET	68.07	64.15	64.80	72.29
AdaBoost	76.54	68.85	73.04	79.23
FFNN	77.02	71.03	75.34	**81.51**

Table 1. Detection accuracy (in %) for four different classifiers with different feature sets. MFS: the mel-frequency cepstral coefficient set, GFS: the glottal feature set. The length of the feature vector is given in parenthesis.

5.2 Narendra, Alku: Glottal source information for pathological voice detection. IEEE Access 8(1): 67745–67755, 2020.

- Binary detection experiments: dysarthric *vs*. healthy (the UA-Speech database and the TORGO database) and dysphonic vs. healthy voices (the UPM database).
- Note: dysarthria in this article is due to ALS and CP. For our similar studies on Parkinson's disease, see Liu *et al*., 2023 & Narendra, Schuller and Alku, 2021.
- Study goal: to investigate the use of the estimated glottal flow in detection of pathological voice by comparing the classical pipeline approach to the end-to-end approach.
- System architectures:

    - a classical pipeline system (with the SVM classifiers and different features)

    - an end-to-end system (using either the raw voice signal or the estimated glottal flow)

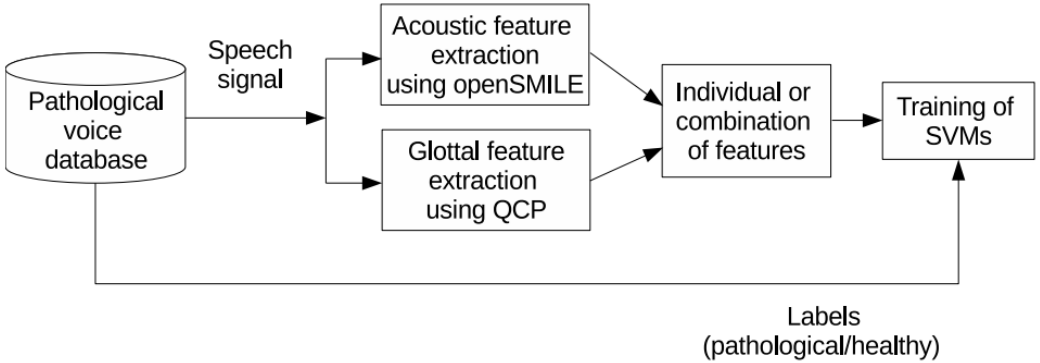


Fig. 13. The classical pipeline system (training stage).

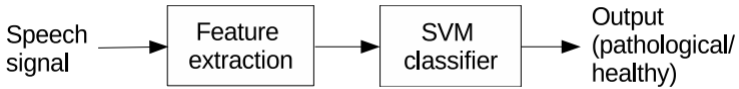


Fig. 14. The classical pipeline system (test stage).

	UA-Speech			TORGO			UPM		
Feature set	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
openSMILE-1	76.65	73.76	79.54	78.24	72.94	83.54	70.73	78.49	62.98
openSMILE-2	86.99	82.35	91.63	80.62	73.73	87.52	72.50	77.61	67.84
Glottal-1	73.56	76.47	70.65	67.17	71.22	63.12	63.17	67.08	59.27
Glottal-2	68.74	69.36	68.12	66.93	71.55	62.17	64.63	65.85	63.41
openSMILE-1 + Glottal-1	81.01	82.55	79.48	79.62	73.21	86.03	73.39	79.54	67.24
openSMILE-2 + Glottal-1	**91.88**	**92.56**	**91.21**	**82.12**	**79.02**	85.22	75.61	79.11	72.20
openSMILE-1 + Glottal-2	80.02	81.93	78.12	80.63	72.59	**88.68**	74.17	80.05	68.29
openSMILE-2 + Glottal-2	91.19	91.57	90.82	81.35	76.83	85.87	**76.83**	**80.81**	**73.36**

Table 2. Classification results obtained using the classical pipeline systems with the openSMILE and glottal features for the three pathological voice databases

Reading instructions:

- openSMILE-1 and openSMILE-2: widely used acoustic feature sets (no inverse filtering)
- Glottal-1 and Glottal-2: glottal feature sets (based on inverse filtering)
- UA-Speech, TORGO, UPM: databases of pathological speech
- accuracy, sensitivity, specificity: widely used metrics (optimal value: 100%)

6. Summary and conclusions

- Glottal inverse filtering (GIF) enables non-invasive estimation of the voice source from microphone speech signals.
- Many GIF methods and glottal parameters have been developed during past decades, and new GIF methods have been proposed recently.
- The QCP method (Airaksinen *et al*., 2017) is available as MATLAB code and as implementation in the Aalto Aparat GIF tool (Alku *et al*., 2017).
- GIF has been used in fundamental research, but the topic was handled in the current presentation mainly from the point of view speech-based biomarking of health that uses machine learning (ML) classifiers.
- GIF can be used with ML in biomarking of health both in (1) classical pipeline classifiers (in feature extraction) and (2) in end-to-end classifiers (as an alternative input to the speech signal).

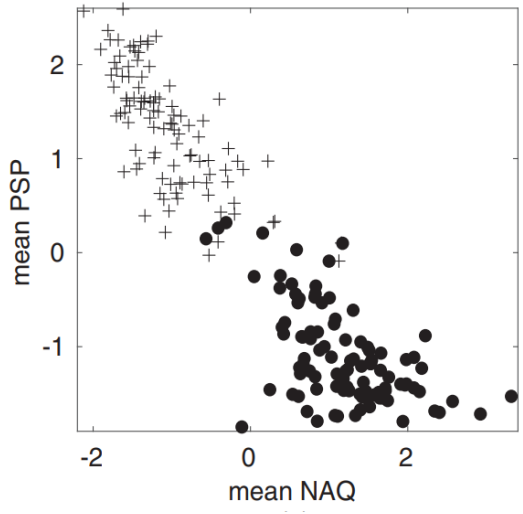
- Adding glottal features to widely used speech features (*e.g*. MFCCs, openSMILE) has been shown to improve detection accuracy in several recent biomarking studies.

  => Glottal features *per se* are not necessarily best features for biomarking but they include complementary information about phonation that has been shown to improve existing speech features.

- In speech-based biomarking of health, deep learning –based end-to-end approaches are popular today. Therefore, the computation of glottal (and other) features might sound less attractive. However, the following issues justify the use of glottal features.

  (1) Glottal features provide better explainability compared to deep learning –based end-to-end systems (“black boxes”).

  (2) Since voice source information is embedded in the speech signal, this information should be able to be extracted by non-linear end-to-end networks without using GIF. However, end-to-end networks are data hungry and they cannot necessarily be trained properly in the area of pathological voice, where little training data is typically available.



Glottal parameters used:

NAQ (normalized amplitude quotient, Alku *et al*., 2002): a measure of relative glottal closing phase length. Large value: breathy phonation, small value: pressed phonation.

PSP (parabolic spectral parameter, Alku *et al.*, 1997): a measure of spectral tilt. Large value: small tilt, small value: large tilt.

Fig. 15. An example demonstrating dysarthric speech of two severity levels. The speech signals are expressed in a 2-dim space spanned by two glottal parameters for low disease severity (filled circles) and for high disease severity ('+' marks). (Narendra and Alku, 2020). Both parameters were z-score normalized (mean=0; std=1).

- Out latest biomarking studies have addressed other than GIF-related topics such as exemplar-based sparse representations in detection of Parkinson's disease (Reddy and Alku, 2023), wavelet scattering features (*e.g.* Keerthana *et al.*, 2024) and pre-trained models (*e.g*. Tirronen, Kadiri and Alku, 2023; Javanmardi, Kadiri and Alku, 2024).

References

Airaksinen, Bäckström, Alku (2017). Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization, IEEE Trans. Audio Speech Lang. Process. 25(4): 929–939.

Airaksinen, Raitio, Story, Alku (2014). Quasi closed phase glottal inverse filtering analysis with weighted linear prediction, IEEE Trans. Audio Speech Lang. Process. 22(3): 596–607.

Alku (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, Speech Commun. 11(2–3): 109–118.

Alku, Bäckström, Vilkman (2002). Normalized amplitude quotient for parameterization of the glottal flow, J. Acoust. Soc. Am. 112(1): 701–710.

Alku, Pohjalainen, Airaksinen (2017). Aalto Aparat – A freely available tool for glottal inverse filtering and voice source parameterization, Proc. Subsidia: Tools and Resources for Speech Sciences.

Alku, Strik, Vilkman (1997). Parabolic spectral parameter – A new method for quantification of the glottal flow, Speech Commun. 22: 67–79.

Alzamendi, Schlotthauer (2017). Modeling and joint estimation of glottal source and vocal tract filter by state-space methods, Biomed. Signal Process. Control 37: 5–15.

Bozkurt, Doval, d'Alessandro, Dutoit (2005). Zeros of z-transform representation with application to source-filter separation in speech, IEEE Signal Process. Lett. 12(4): 344–347.

Childers, Lee (1991). Vocal quality factors: Analysis, synthesis, and perception, J. Acoust. Soc. Am. 90(5): 2394–2410.

Cortes *et al*. (2022). Kalman filter implementation of subglottal impedance-based inverse filtering to estimate glottal airflow during phonation, Appl. Sci. 12: article no. 401.

Cortes, Vapnik (1995). Support-vector networks, Machine Learning 20 (3): 273–297.

Davis, Mermelstein (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28: 357–366.

Deshmukh, Al Ismail, Singh (2021). Interpreting glottal flow dynamics for detecting Covid-19 from voice, Proc. Icassp, pp. 1055–1059.

Drugman, Bozkurt, Dutoit (2011). Causal–anticausal decomposition of speech using complex cepstrum for glottal source estimation, Speech Commun. 53: 855–866.

Eyben, Wöllmer, Schuller (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor, Proc. ACM Int. Conf. on Multimedia, pp. 1459–1462.

Fant (1961). A new anti-resonance circuit for inverse filtering, Speech Transmission Laboratory Quarterly Progress and Status Report 2(1):1–6.

Fröhlich, Michaelis, Strube (2001). SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals, J. Acoust. Soc. Am. 110(1): 479–488.

Fu, Murphy (2006). Robust glottal source estimation based on joint source-filter model optimization, IEEE Trans. Audio Speech Lang. Process. 14(1): 492–501.

Gomez-Vilda *et al*. (2009). Glottal source biometrical signature for voice pathology detection, Speech Commun. 51: 759–781, 2009.

Javanmardi, Kadiri, Alku (2024). Pre-trained models for detection and severity level classification of dysarthria from speech, Speech Commun. 158: article no. 103047.

Kadiri, Alku (2020). Analysis and detection of pathological voice using glottal source features, IEEE J. Sel. Top. Signal Process. 14(2): 367–379.

Kaushik, Baghel, Burget, Travieso, Dutta (2021). SLINet: Dysphasia detection in children using deep neural network, Biomed. Signal Process. Control 68: article no. 102798.

Keerthana, Reddy, Alku, Rao, Mitra (2024). Automatic classification of neurological voice disorders using wavelet scattering features, Speech Commun. 157: article no. 103040.

Langheinrich, Stone, Zhang, Birkholz (2022). Glottal inverse filtering based on articulatory synthesis and deep learning, Proc. Interspeech, pp. 1327–1331.

Lindqvist-Gauffin (1965). Studies of the voice source by means of inverse filtering, Speech Transmission Laboratory Quarterly Progress and Status Report 6(1): 8–13

Liu *et al*. (2023). Automatic assessment of Parkinson's disease using speech representations of phonation and articulation, IEEE Trans. Audio Speech Lang. Process. 31: 242–255.

Ma, Kamp, Willems (1993). Robust signal selection for linear prediction analysis of voiced speech, Speech Commun. 12(1): 69–81.

Miller (1959). Nature of the vocal cord wave, J. Acoust. Soc. Am. 31(6): 667–677.

Mokhtari, Story, Alku, Ando (2018). Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production, Speech Commun. 104: 24–38.

Monsen, Engebretson (1977). Study of variations in the male and female glottal wave, J. Acoust. Soc.Am. 62(4): 981–993.

Nakatsui, Suzuki (1970). Method of observation of glottal-source wave using digital inverse filtering in time domain, J. Acoust. Soc. Am. 47(1): 664–665.

Narendra, Airaksinen, Story, Alku (2019). Estimation of the glottal source from coded telephone speech using deep neural networks, Speech Commun. 106: 95–104.

Narendra, Alku (2020). Automatic intelligibility assessment of dysarthric speech using glottal parameters, Speech Commun.123: 1–9.

Narendra, Schuller, Alku (2021). The detection of Parkinson's disease from speech using voice source information, IEEE Trans. Audio Speech Lang. Process. 29: 1925–1936.

Novotny *et al*. (2020). Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance, Biomed. Signal Process. Control 57: article no. 101818.

Ooi *et al*. (2012). Early prediction of major depression in adolescents using glottal wave characteristics and Teager energy parameters, Proc. Icassp, pp. 4613–4616.

Oppenheim, Schafer (1968). Homomorphic analysis of speech, *IEEE Trans. Audio Electroacoust.* 16(1): 221–226.

Reddy, Alku (2023). Exemplar-based sparse representations for detection of Parkinson's disease from speech, IEEE Trans. Audio Speech Lang. Process. 31: 1386-1396.

Rothenberg (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing, J. Acoust. Soc. Am. 53(6): 1632–1645.

Sahoo, Routray (2016). A novel method of glottal inverse filtering, IEEE Trans. Audio Speech Lang. Process. 24(7): 1230–1241.

Sapienza, Stathopoulos, Dromey (1998). Approximations of open quotient and speed quotient from glottal airflow and EGG waveforms: Effects of measurement criteria and sound pressure level, J. Voice 12(1): 31–43.

Simantiraki *et al*. (2017). Glottal source features for automatic speech-based depression assessment, Proc. Interspeech, pp. 2700–2704.

Strik, Boves (1992). On the relation between voice source parameters and prosodic features in connected speech, Speech Commun. 11: 167–174.

Strube (1974). Determination of the instant of glottal closure from the speech wave, J. Acoust. Soc. Am. 56(5): 1625–1629.

Sundberg (2017). Flow glottogram and subglottal pressure relationship in singers and untrained voices, J. Voice 32(1): 23–31.

Sundberg, Fahlstedt, Morell (2005). Effects on the glottal voice source of vocal loudness variation in untrained female and male voices, J. Acoust. Soc. Am. 117(2): 879–885.

Timcke, von Leden, Moore (1958). Laryngeal vibrations: measurements of the glottic wave, Archiv.Otolaryngol. 68: 1–19.

Tirronen, Kadiri, Alku (2023). Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features, IEEE Open J. Signal Process. 4: 80–88.

Titze, Sundberg (1992). Vocal intensity in speakers and singers, J. Acoust. Soc. Am. 91(5): 2936–2946.

Szklanny, Wrzeciono (2019). The application of a genetic algorithm in the noninvasive assessment of vocal nodules in children, IEEE Access 7: 44966–44976.

Vasquez-Correa *et al*. (2021). On modeling glottal source information for phonation assessment in Parkinson's disease, Proc. Interspeech, pp. 26–30.

Wong, Markel, Gray (1979). Least squares glottal inverse filtering from the acoustic speech waveform, *IEEE Trans. Acoust. Speech Signal Process.* 27: 350–355.

Wu *et al*. (2020). Investigation and evaluation of glottal flow waveform for voice pathology detection, IEEE Access 9: 30–44.