Classification-Based Detection of Glottal Closure Instants from Speech Signals

Jindřich Matoušek and Daniel Tihelka

Dept. of Cybernetics, NTIS – New Technologies for the Information Society
Faculty of Applied Sciences, University of West Bohemia, Czech Republic

jmatouse@kky.zcu.cz, dtihelka@ntis.zcu.cz

1. Introduction

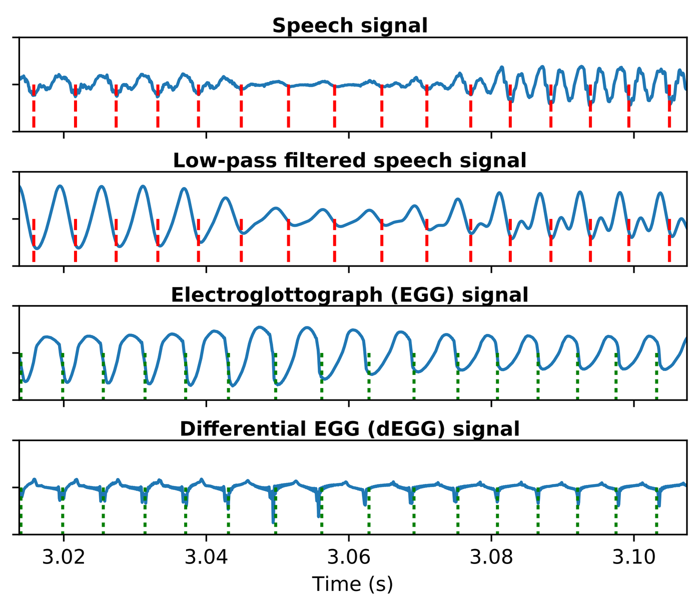
Glottal Closure Instants (GCIs)

- used for pitch-synchronous processing
- defined as speech signal amplitude extreme that corresponds to the moment of glottal closure
- ➠ precise GCI detection important in many speech-technology applications
- various algorithms proposed to detect GCIs directly in the speech signal [1]
- manual tuning often required

Problem Definition

- based on a classification framework
- a classifier trained on relevant features extracted around potential GCI locations (peaks in speech waveform)
- ➠ GCI detection viewed as a two-class classification problem: whether or not a peak represents a GCI
- best performance with these classifiers:
  - extremely randomized trees (ERT)
  - support vector machines (SVM with RBF kernel)
  - k-nearest neighbors (KNN)
  - multilayer perceptron (MLP)

Signals Used for Detection



- EGG signals used for reliable detection but:
  - not always available (only speech often recorded)
  - uncomfortable to record EGG signal
- ➠ GCI detection directly from the speech signal is very important

Aim of this Study

- ➠ to propose a speech-only-based high-quality data-based GCI detection method with the parameters being set up automatically

2. Experimental Data & Features

Data Description

- in-house clean speech data (UWB)
- primarily intended for speech synthesis
- various speakers and languages included:
  - Czech (male and female)
  - Slovak (female)
  - German (male)
  - US English (male)
  - French (female)
- true GCIs produced by a human expert

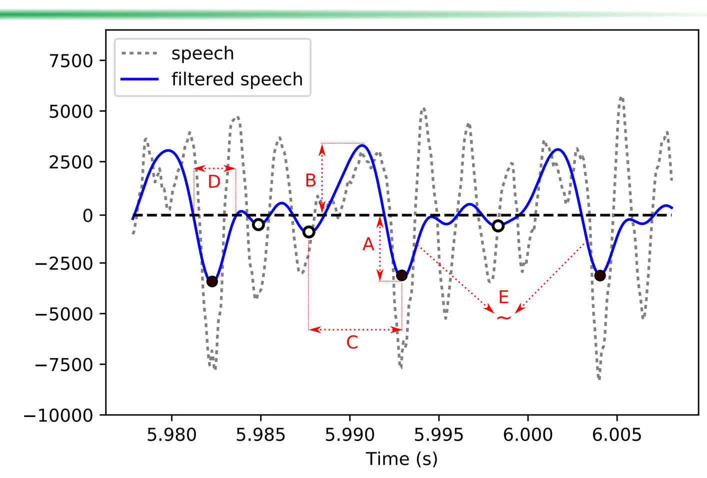
	Training	Validation	Total
# utterances	63	19	82
speech duration	9 min	3 min	12 min
GCI candidates	66,130	18,026	84,156
True GCIs	40,938	10,691	51,629

Speech Signal Pre-Processing

- speech signal low-pass filtered to reduce the high-frequency structure [2]
  - zero-phase Equiripple-designed filter
  - 0.5 dB ripple in the pass band
  - 60 dB attenuation in the stop band
  - 700 Hz cutoff frequency
- speech signal switched to have negative polarity
- peaks identified by zero-crossing used for feature extraction
- negative peaks taken as candidates for GCI placement
- true GCIs assigned to a corresponding negative peak

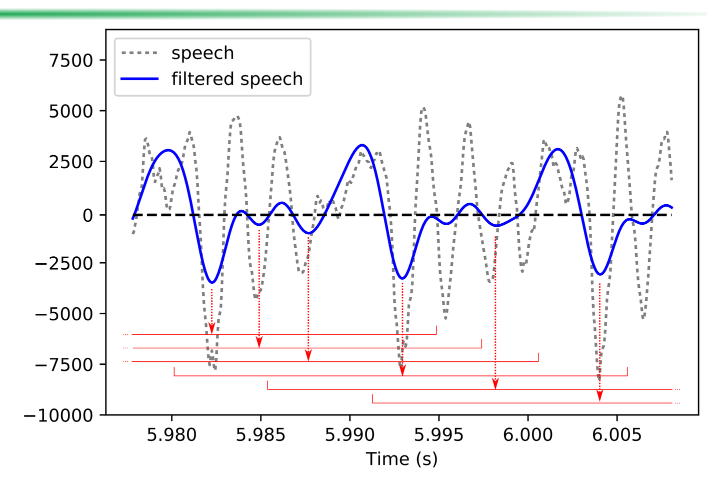
Peak-Based Features

- each negative peak described by a set of local descriptors reflecting the position and shape of other $2P$ neighboring peaks [2]
- $P = 3 \Rightarrow 32$ features in total
  - A: amplitudes of negative peaks (7 features)
  - B: amplitudes of positive peaks (6)
  - C: time difference between negative peaks (6)
  - D: width of negative peaks (7)
  - E: correlation of negative peaks (6)



Waveform Sample-Based Features

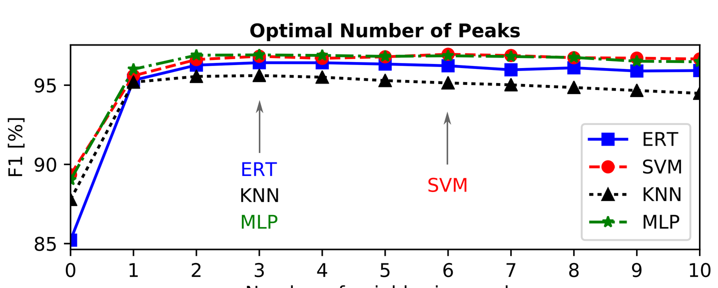
- hanning-windowed waveform samples around a negative peak
- for window length 30 ms ($S = 30$) and frequency sampling 16 kHz:
  - the current sample of a negative peak
  - 240 preceding samples
  - 240 succeeding samples
  - ➠ 481 features in total

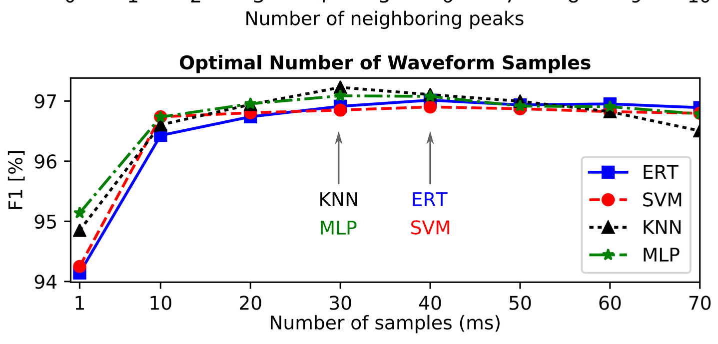


3. Classifier Design

Feature Engineering

- search for optimal number of features
  - number of peaks surrounding each negative peak
  - number of samples around each negative peak
  - default classifier hyper-parameters according to Scikit-learn toolkit [3]
  - 10-fold cross validation on the training dataset

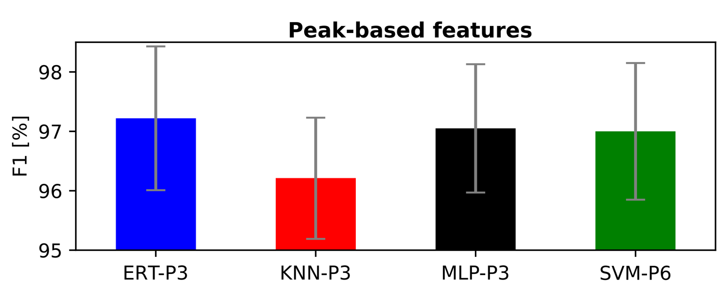


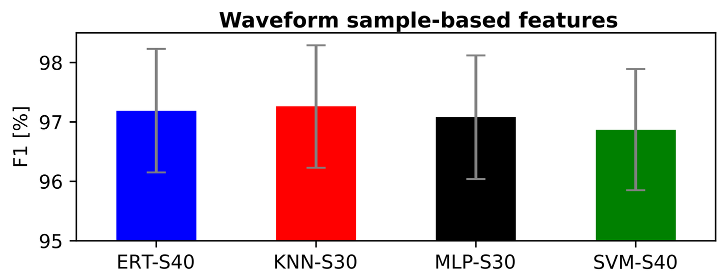


Hyper-Parameter Tuning

- grid search over relevant values of classifier hyper-parameters with the optimal features
- 10-fold cross validation on the training dataset
- evaluation in terms of $F1$-score and 95% confidence intervals

CLF-Pm … No. of peaks prior and subsequent to current peak for classifier CLF
CLF-Sn … window length (ms) around the current peak for classifier CLF





4. Classifier Selection & Evaluation

Classification-Based GCI Detection Results

- performed on UWB validation dataset
- best classifiers for both kind of features selected (ERT-P3 and KNN-S30)
- combination of both kinds of features (ERT-P3S30 and KNN-P3S30) also evaluated
- Recall (R), Precision (P), and $F1$-score used

Classifier	R (%)	P (%)	$F1$ (%)
ERT-P3	96.46	98.09	97.27
KNN-S30	96.45	97.75	97.10
ERT-P3S30	96.74	97.65	97.20
KNN-P3S30	96.68	97.80	97.23

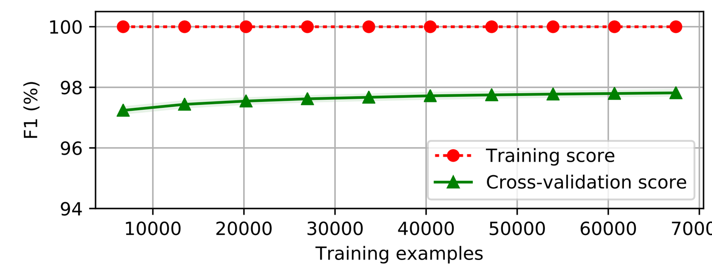
Statistical Significance

- McNemar's test
  - $\gg$ significantly better ($\alpha = 0.01$)
  - $>$ significantly better ($\alpha = 0.05$)
- ➠ ERT-P3 performs significantly better than other classifiers

Classifier	ERT-P3	KNN-S30	ERT-P3S30	KNN-P3S30
ERT-P3	=	$\gg$	$\gg$	$\gg$
KNN-S30	$\ll$	=	<	<
ERT-P3S30	$\ll$	>	=	=
KNN-P3S30	$\ll$	>	=	=

Learning Curves

- learning curves for ERT-P3 classifier
- still some room for improvement
  - more training data
  - other features



5. Comparison with Other Methods

Methods

- ERT-P3
  github.com/ARTIC-TTS-experiments/2017_Interspeech
- SEDREAMS (COVAREP repository) [4]
  github.com/covarep
- Microcanonical Multiscale Formalism (MMF) [5]
  geostat.bordeaux.inria.fr/index.php/downloads.html
- DYPSA (VOICEBOX toolbox) [6]
  www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
- detected GCIs filtered by voiced/unvoiced detector (RAPT) and shifted towards the neighboring negative peak

Datasets

Dataset	Lang.&sex	# utts	Mins
UWB	see Sec. 2	19	3
BDL (CMU ARCTIC)	US male	1132	54
SLT (CMU ARCTIC)	US female	1132	54
KED (CSTR TIMIT)	US male	453	20

- UWB: hand-crafted reference GCIs used
- ARCTIC, TIMIT: no hand-crafted GCIs available $\Rightarrow$ reference GCIs detected from EGG recordings using MPA [7]
- MPA also used as upper bound for UWB

Results

Dataset	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	A25 (%)	E10 (%)
UWB	MPA	97.06	0.66	2.28	0.21	84.65	97.03
	ERT-P3	95.87	1.99	2.14	0.29	81.06	95.93
	SEDREAMS	91.80	3.54	4.66	0.24	81.51	91.87
	MMF	83.47	11.42	5.11	0.42	80.72	84.80
	DYPSA	87.40	4.86	7.74	0.40	80.60	87.27
BDL	ERT-P3	91.96	2.98	5.06	0.41	88.41	91.78
	SEDREAMS	90.98	2.35	6.67	0.54	91.23	90.57
	MMF	87.82	5.84	6.34	0.61	90.36	87.77
	DYPSA	86.98	7.59	5.43	0.65	91.16	86.69
SLT	ERT-P3	95.18	1.35	3.47	0.15	95.08	95.07
	SEDREAMS	92.96	1.15	5.89	0.19	89.09	92.61
	MMF	91.16	5.33	3.51	0.37	77.53	91.32
	DYPSA	91.50	2.80	5.70	0.30	81.23	91.24
KED	ERT-P3	91.88	2.94	5.18	0.27	88.02	91.69
	SEDREAMS	89.54	1.16	9.30	0.56	78.46	88.61
	MMF	89.11	4.61	6.28	0.57	83.52	88.92
	DYPSA	89.01	4.62	6.37	0.48	83.70	88.81

Reliability:

$$MR = \frac{N_M}{N_R}$$
$$FAR = \frac{N_{FA}}{N_R}$$
$$IDR = 1 - MR - FAR$$

Accuracy:

$$A25 = \frac{N_{\zeta \le 0.25}}{N_R - N_M - N_{FA}}$$
$$IDA = \mathrm{stdev}(\zeta)$$

Combined dynamic measure:

$$E10 = \frac{N_R - N_{\zeta > 0.1T_0} - N_M - N_{FA}}{N_R}$$

N_R … # reference GCIs
N_M … # missing GCIs
N_{FA} … # false GCIs
ζ … identification error of corresp. GCIs
$N_{\zeta \le 0.25}$ … # corresp. GCIs with $\zeta \le 0.25$ ms
$N_{\zeta > 0.1T_0}$ … # corresp. GCIs with $\zeta > 0.1T_0$

6. Conclusions

Conclusion

- classification-based GCI detection proposed
- data-based method $\Rightarrow$ only true GCIs required; classifier parameters trained automatically
- the proposed method outperformed other state-of-the-art methods on several test datasets in terms of detection reliability and mostly also in terms of accuracy

Future work

- performance on more data from more speakers
- incorporation of other features (pitch-based, voiced/unvoiced or harmonic/noise related)
- only clean speech data investigated so far $\Rightarrow$ performance on noisy signals and emotional/expressive speech?
- deep learning?

References

[1] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, & T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, 2012.

[2] E. Barnard, A. Cole, M.P. Vea, & F.A. Alleva, "Pitch detection with a neural-net classifier," *IEEE Trans. Signal Process.*, vol. 39, no. 2, 1991.

[3] F. Pedregosa, G. Varoquaux, et. al, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol 12, 2011.

[4] T. Drugman & T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *INTERSPEECH*, 2009.

[5] V. Khanagha, K. Daoudi, & H.M. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, 2014.

[6] P.A. Naylor, A. Kounoudes, J. Gudnason, & M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, 2007.

[7] M. Legát, J. Matoušek, & D. Tihelka, "A robust multi-phase pitch-mark detection algorithm," in *INTERSPEECH*, 2007.

This research was supported by the Czech Science Foundation (GA CR), project No. GA16-04420S. The access to the MetaCentrum clusters provided under the programme LM2010005 is highly appreciated.