A LANDMARK-BASED APPROACH TO AUTOMATIC VOICE ONSET TIME ESTIMATION IN STOP-VOWEL SEQUENCES

Stephan R. Kuberski Stephen J. Tobin Adamantios I. Gafos

University of Potsdam
Linguistics Department
Potsdam, Germany

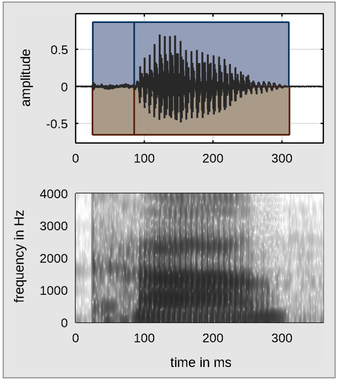
IEEE GlobalSIP, December 7–9, 2016

Outline

Terminology

- **Example:** stop-vowel sequence /ka/, German male speaker, age: 24

- **Voice onset time (VOT):** length of the interval between the release of an oral closure and the onset of vocal fold vibrations
- **Release burst:** abrupt increase in acoustic energy caused by release of constriction of plosive consonants (e.g., /t/, /k/, /p/)
- **Voicing:** presence of vocal fold vibrations during the production of speech sounds (e.g., voiced stops: /d/, /g/, /b/)
- voicing is typically present during production of German vowels (glottal activity)
- plosive consonants with different place of articulation (e.g., /t/ versus /k/) differ in VOT values (**linguistic contrast**)



Estimation system

Implicit systems

- usually statistical learning methods
- supervised learning requires a subset of previously (manually) labeled data
- often no explicit output of utilized delimiting landmarks

Explicit systems

- usually knowledge-/rule-based expert systems
- no need of previously labeled data
- explicit output of delimiting landmarks

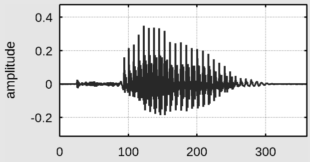
Proposed approach

- **explicit landmark detection** of release burst (+b), glottal activity onset (+g) and offset (–g)
- subsequent application of a **set of rules** to verify candidate landmarks

Release burst detection (Ananthapadmanabha et al., 2014)

1) use equal loudness filtered signal $x[n]$

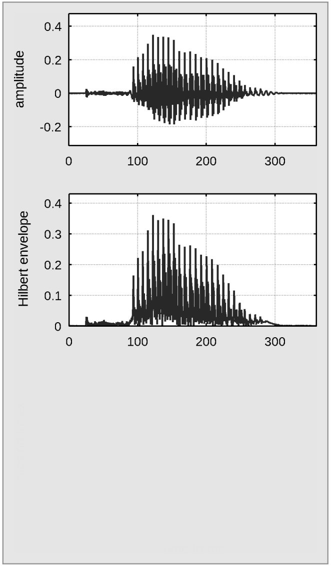
2) consider subsets between zero crossings $n_1, n_2, \dots$



Release burst detection (Ananthapadmanabha et al., 2014)

1) use equal loudness filtered signal $x[n]$
2) consider subsets between zero crossings $n_1, n_2, \dots$
3) compute discrete Hilbert envelope

$$H[n] = \left| x[n] + \frac{\mathrm{i}}{\pi} \sum_{\substack{k=-\infty \\ k \neq n}}^{\infty} \frac{x[k]}{n-k} \right|$$



Release burst detection (Ananthapadmanabha et al., 2014)

1) use equal loudness filtered signal $x[n]$
2) consider subsets between zero crossings $n_1, n_2, \ldots$
3) compute discrete Hilbert envelope

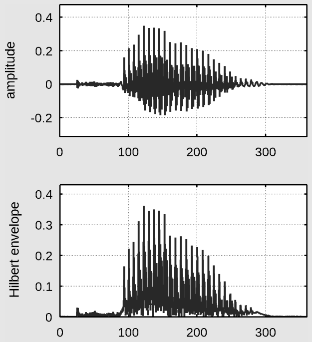
$$H[n] = \left| x[n] + \frac{\mathrm{i}}{\pi} \sum_{\substack{k=-\infty \\ k \neq n}}^{\infty} \frac{x[k]}{n-k} \right|$$

4) for each subset compute maximum Hilbert envelope

$$m_{i,\max} = \underset{n_i \leq m \leq n_{i+1}}{\arg\max}\ H[m], \qquad H_{i,\max} = H[m_{i,\max}]$$

5) set average of preceding vicinity $[m_{i,1}, m_{i,2}]$ (10 ms + 1 ms)

$$H_{i,\text{avg}} = \frac{1}{m_{i,2} - m_{i,1} + 1} \sum_{k=m_{i,1}}^{m_{i,2}} H[k]$$



Release burst detection (Ananthapadmanabha et al., 2014)

1) use equal loudness filtered signal $x[n]$
2) consider subsets between zero crossings $n_1, n_2, \ldots$
3) compute discrete Hilbert envelope

$$H[n] = \left| x[n] + \frac{\mathrm{i}}{\pi} \sum_{\substack{k=-\infty \\ k \neq n}}^{\infty} \frac{x[k]}{n-k} \right|$$

4) for each subset compute maximum Hilbert envelope

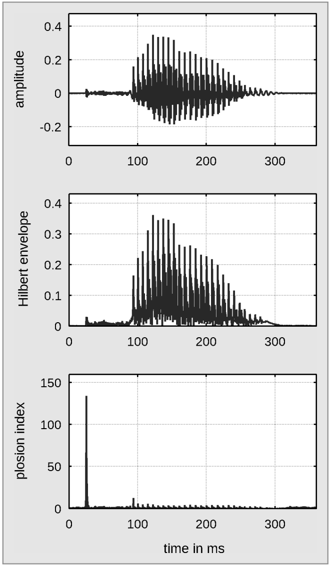
$$m_{i,\max} = \underset{n_i \leq m \leq n_{i+1}}{\arg\max}\, H[m], \qquad H_{i,\max} = H[m_{i,\max}]$$

5) set average of preceding vicinity $[m_{i,1}, m_{i,2}]$ (10 ms + 1 ms)

$$H_{i,\text{avg}} = \frac{1}{m_{i,2} - m_{i,1} + 1} \sum_{k=m_{i,1}}^{m_{i,2}} H[k]$$

6) define plosion index at vicinity onset

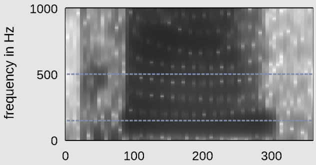
$$I[n = m_{i,1}] = \frac{H_{i,\max}}{H_{i,\text{avg}}}, \qquad I[n > m_{i,1}] = 0$$



Glottal activity detection (Liu, 1996)

1) use signal's short time Fourier transform (15 ms window)

$$X[m,\omega] = \sum_{k=-\infty}^{\infty} w[k-m]x[k]\,\mathrm{e}^{-\mathrm{i}\omega k}$$



Glottal activity detection (Liu, 1996)

1) use signal's short time Fourier transform (15 ms window)

$$X[m,\omega] = \sum_{k=-\infty}^{\infty} w[k-m]x[k]e^{-i\omega k}$$

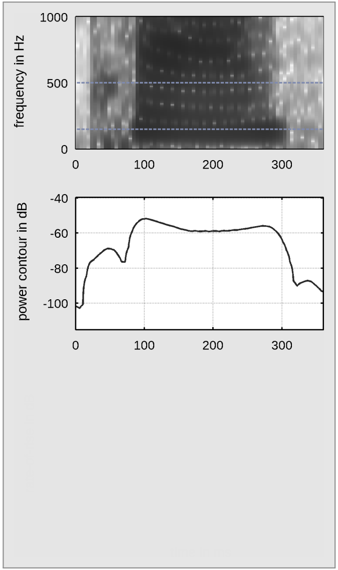
2) compute subband (150...500 Hz) power contour

$$P[m] = \max_{\omega_{\min} \le \omega \le \omega_{\max}} |X[m,\omega]|^2$$

3) undo short time segmentation: $P[m] \rightsquigarrow P[n]$

4) apply box blur kernel (20 ms width)

$$P[n] = \sum_{l=1}^{2L} k[l]P[n+l-L]$$



Glottal activity detection (Liu, 1996)

1) use signal's short time Fourier transform (15 ms window)

$$X[m,\omega] = \sum_{k=-\infty}^{\infty} w[k-m]x[k]\mathrm{e}^{-\mathrm{i}\omega k}$$

2) compute subband (150...500 Hz) power contour

$$P[m] = \max_{\omega_{\mathrm{min}} \leq \omega \leq \omega_{\mathrm{max}}} |X[m,\omega]|^2$$

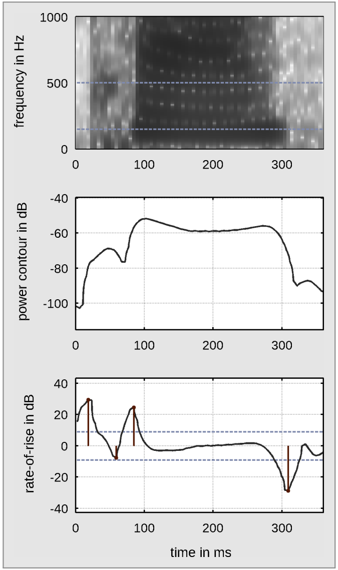
3) undo short time segmentation: $P[m] \rightsquigarrow P[n]$

4) apply box blur kernel (20 ms width)

$$P[n] = \sum_{l=1}^{2L} k[l]P[n+l-L]$$

5) compute power rate-of-rise (12.5 ms lookahead/-behind)

$$R[n] = P[n+w_{\mathrm{a}}] - P[n-w_{\mathrm{b}}]$$



Glottal activity detection (Liu, 1996)

1) use signal's short time Fourier transform (15 ms window)

$$X[m,\omega] = \sum_{k=-\infty}^{\infty} w[k-m]x[k]\mathrm{e}^{-\mathrm{i}\omega k}$$

2) compute subband (150...500 Hz) power contour

$$P[m] = \max_{\omega_{\min} \leq \omega \leq \omega_{\max}} |X[m,\omega]|^2$$

3) undo short time segmentation: $P[m] \rightsquigarrow P[n]$

4) apply box blur kernel (20 ms width)

$$P[n] = \sum_{l=1}^{2L} k[l]P[n+l-L]$$

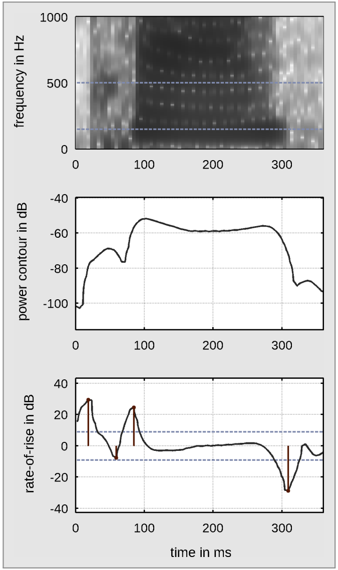
5) compute power rate-of-rise (12.5 ms lookahead/-behind)

$$R[n] = P[n+w_\mathrm{a}] - P[n-w_\mathrm{b}]$$

6) detect ±peaks exceeding a certain threshold (±9 dB)

7) ensure natural peak pairing using insertions and deletions

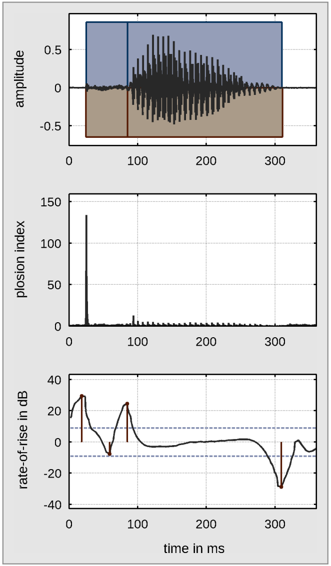
8) no leading –peak, no trailing +peak



Voice onset time estimation

- verify candidate landmarks of release burst (+b), voice onset (+g) and voice offset (–g) by means of **additional rules**:

1) any (±g) pair located completely in the first third is discarded (consonant to vowel transition)
2) merge remaining successive (±g) pairs into a single pair bypassing any gaps
3) choose most significant plosion index in front of and closest to that single pair

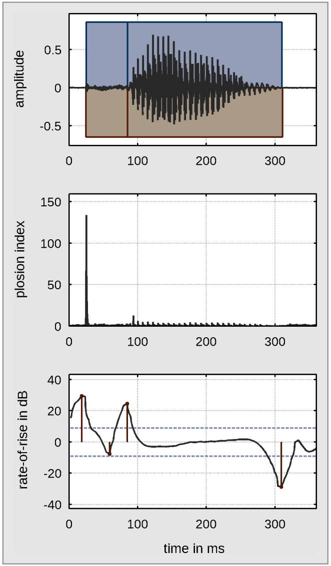


Voice onset time estimation

- verify candidate landmarks of release burst (+b), voice onset (+g) and voice offset (–g) by means of **additional rules**:

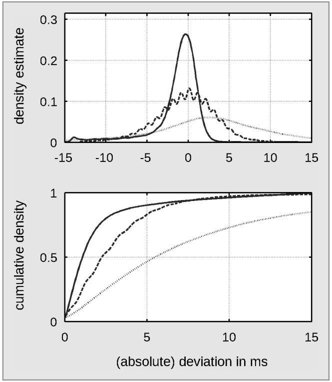
1) any (±g) pair located completely in the first third is discarded (consonant to vowel transition)
2) merge remaining successive (±g) pairs into a single pair bypassing any gaps
3) choose most significant plosion index in front of and closest to that single pair

- yield final landmarks of **release burst** (+b) (step 3) and **voice onset** (+g) (step 2)
- **voice onset time (VOT)** is the length of the interval between those two landmarks
- additional **voice offset** (–g) landmark is available (e.g., useful for VOT normalization by syllable length)



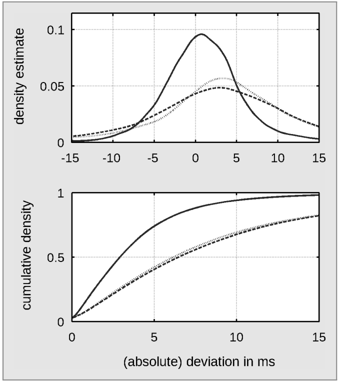
Results (1)

Landmark detection



Landmark	5 ms	10 ms	15 ms
burst onset (+b)	90.4	96.1	99.6
voice onset (+g)	83.0	97.1	98.6
voice offset (–g)	46.5	72.9	85.0

Interval estimation



Interval	5 ms	10 ms	15 ms
voice onset time	73.9	94.0	98.1
vowel length	40.3	67.6	82.0
syllable length	42.2	69.3	82.5

Results (2)

Our dataset

- registered for the purposes of experiments described in Klein et al. (2015)
- clean acoustic speech recordings (sound booth, 16 bit mono, 44100 Hz)
- 42 native German speakers (29 female, 13 male, aged between 18 and 44)
- 40021 isolated stop-vowel tokens (19881 /ka/, 20140 /ta/)

TIMIT (subset)

- 168 native American English speakers
- 5459 word-medial stops
- large number of consonant-vowel combinations

Author (and technique)	Accuracy
Stouten and Hamme, 2009 (reassignment spectra)	76.1%
Lin and Wang, 2011 (random forests)	83.4%
Sonderegger and Keshet, 2012 (structured prediction)	87.6%
Ryant et al., 2013 (support vector machines)	91.7%
proposed approach	94.0%

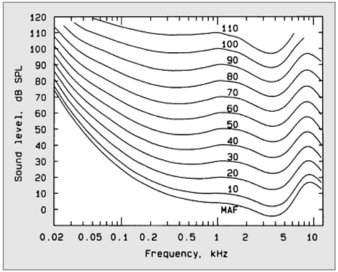
References

- Ananthapadmanabha, T. V., A. P. Pratosh, and A. G. Krishnan (2014). "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index". In: *J. Acoust. Soc. Am.* 135 (1), pp. 460–471. DOI: 10.1121/1.4836055.
- Liu, S. A. (1996). "Landmark detection for distinctive feature-based speech recognition". In: *J. Acoust. Soc. Am.* 100 (5), pp. 3417–3430. DOI: 10.1121/1.416983.
- Klein, E., K. D. Roon, and A. I. Gafos (2015). "Perceptuo-motor interactions across and within phonemic categories". In: *Proc. 18th Int. Congr. Phon. Sci.* Glasgow.
- Stouten, V. and H. van Hamme (2009). "Automatic voice onset time estimation from reassignment spectra". In: *Speech Comm.* 51 (12), pp. 1194–1205. DOI: 10.1016/j.specom.2009.06.003.
- Lin, C. Y. and H. C. Wang (2011). "Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection". In: *J. Acoust. Soc. Am.* 130 (1), pp. 514–525. DOI: 10.1121/1.3592233.
- Sonderegger, M. and J. Keshet (2012). "Automatic measurement of voice onset time using discriminative structured prediction". In: *J. Acoust. Soc. Am.* 132 (6), pp. 3965–3979. DOI: 10.1121/1.4763995.
- Ryant, N., J. Yuan, and M. Liberman (2013). "Automating phonetic measurement: The case of voice onset time". In: *Proc. Mtgs. Acoust.* Vol. 19. Montreal. DOI: 10.1121/1.4801056.

Equal loudness filter (Replay gain)

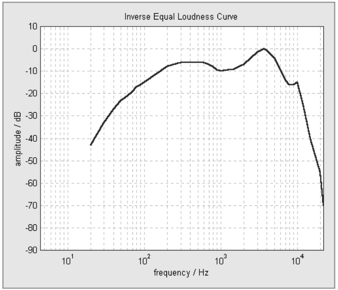
- R. Robinson (2001). *Replay Gain—A Proposed Standard*. http://replaygain.hydrogenaud.io/proposal/equal_loudness.html

Equal loudness curves



- sound pressure required for a test tone of any frequency to sound as loud as a test tone of 1 kHz

Equal loudness filter



- certain benefits over A-, B-, C-, D- and Z-weightings (International standard IEC)