Automatic voice onset time estimation from reassignment spectra (Stouten and Van hamme, 2009)

Cheonkam Jeong

Introduction

- Problem: The settings (i.e., sliding window size and shift) used in the canonical ASR fails to catch acoustic events that occur at a finer time scale, such as Voice Onset Time
  - VOT: time interval between the release of the plosive and the onset of voicing of the following vowel; primary acoustic cue to distinguish plosives in many languages, such as English
  - Some limitations (?) of the traditional ASR: difficult to do modeling of timing at different scales
- Goal: to provide an algorithm that considers phone-level features, such as VOT ⇒ better performance…
- Method: the reassigned time-frequency representation (RTFR), a high resolution signal analysis method

Spectral reassignment

- Time-frequency reassignment (Auger and Flandrin, 1995, among others): to improve the sharpness of the localization of the signal components by reallocating its energy distribution in the time-frequency plane

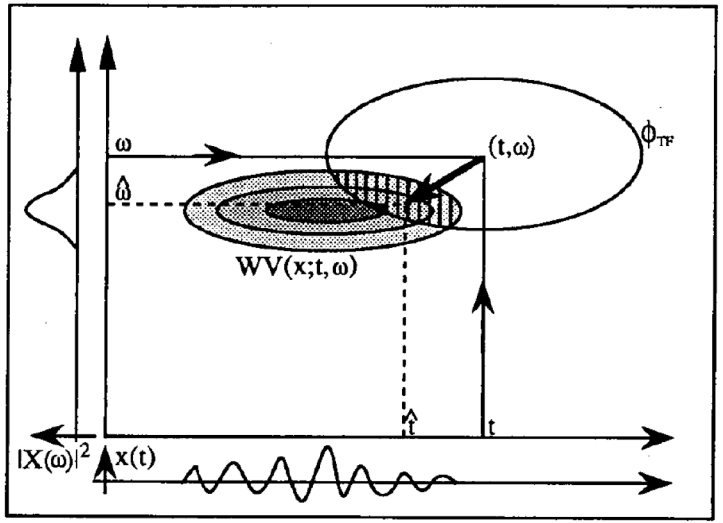


Fig. 1. Principles of the reassignment method.

Spectral reassignment

- 8ms Hamming window, shifted by 0.625 ms per analysis frame, thus 128 and 10 samples, respectively at a sampling frequency of 16 kHz
- 256 equally spaced frequency bins for reassignment

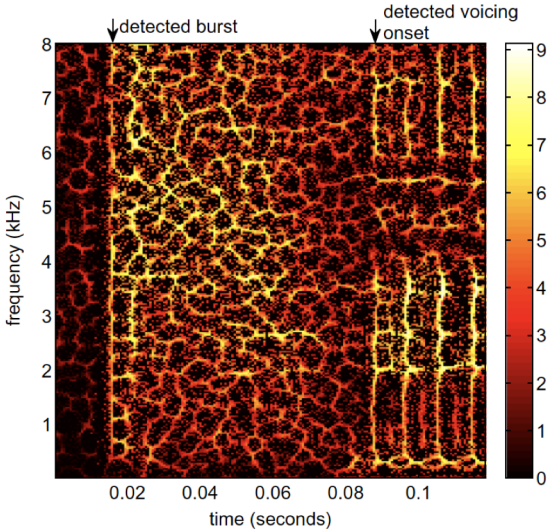


Fig. 1. Reassigned time–frequency representation of a /t/ segment followed by /ih/. Colors encode the logarithm of the energy.

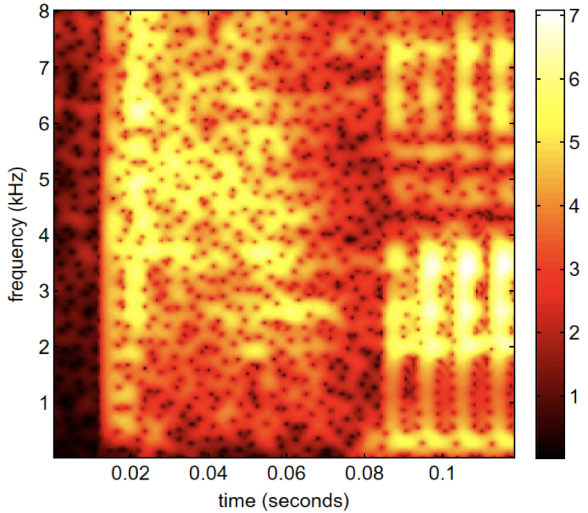


Fig. 2. STFT representation of the /t/ segment from Fig. 1. Color encode the logarithm of the energy.

Properties of the VOT

- Factors that affect VOT values: place of articulation, speech rate, context, position within the word, lexical stress, gender, etc.
- Problem: Voiceless stops & high f0 => shorter VOT, longer VOT for the voiced stops in conversational stops. Due to this overlapping distribution, the VOT value and plosive identity is not straightforward.
- Solution: only consider plosives that are uttered in a constrained way

Data sets

- Data: TIMIT (Garofolo et al., 1990)
- Targets: 6 plosives (/p, t, k, b, d, g/)
- Four data sets: “forced,” “manual,” “free,” and “test”

Table 1
Number of speech segments in each of the data sets.

	Forced	Free	Manual	Test
/b/	2181	2012	115	754
/d/	2432	2222	76	728
/g/	1191	977	98	386
/p/	2588	2749	111	821
/t/	3948	4052	92	1180
/k/	3794	3968	90	1039
Total	16134	15,980	582	4908

Data sets - “forced”

- Segment boundaries using a forced alignment with a HMM-based speech recognizer using the manually verified phonetic transcriptions
- Irrespective of the left and the right phonetic context
- The acoustic models: context independent HMMs with 2-4 states per phone
- The speech features: mel-scaled log-filterbank outputs
- Sharing closure models depending on the voicing of the targets
- Segment boundaries for the plosive: burst only

Data sets - “free”

- Fully automatic VOT extraction setting
- Plosive segment candidates generated by a phonetic automatic speech recognizer as described in (Demuynck et al., 2006) applied to the same utterances used in the “forced” data set
- The HMMs described above used to find the best matching phonetic transcription using a phone-level bigram language model with Witten-Bell smoothing (Witten and Bell, 1991)

Data sets - “manual”

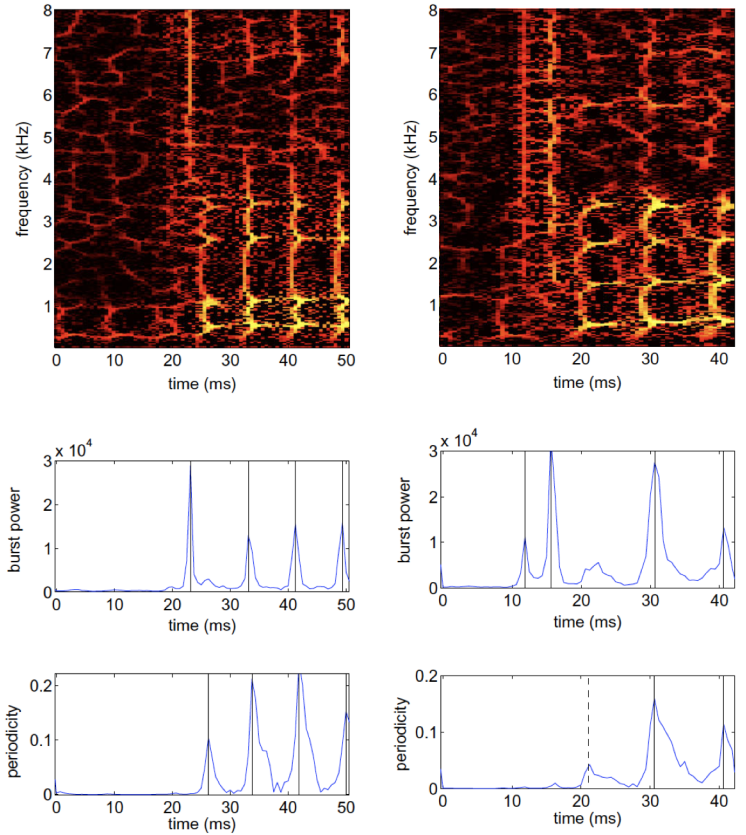
- A subset of the plosive speech segments selected from the “forced” set
- Manually measured by the experts

Data sets - “test”

- constructed exactly like the “forced” data set, except that the sentences are taken from the TIMIT test set

The VOT estimation algorithm

- Step 1: candidate plosive segments are detected and segment boundaries are generated
- Step 2: the burst onset is detected based on “burst power”
- Step 3: the onset of voicing is found based on “periodicity”



The VOT estimation algorithm - detection of plosive segments

- Step 1: candidate plosive segments are detected and segment boundaries are generated
- Using a HMM-based automatic speech recognizer described earlier
- Defined the “forced” and “free” data with/without phonetic knowledge of the test utterance
- The burst 2.5 ms or 4 frames prior to the burst segment start found by the recognizer

The VOT estimation algorithm - burst onset detection

- Step 2: the burst onset is detected based on “burst power”
- The corresponding frequency bins in the RTFR power summed to form the “burst power” p(n)
  - $p(n) > p(n - j)$, for $j = -1, 1$, and 2 (local maximum)
  - $p(n) - p(n - i) > p_m(n)$ for $i = 2, \ldots, 5$ (sufficiently sharp and strong peak), where $p_m(n)$ is taken to be mean of p(n) over 150 plosive frames

The VOT estimation algorithm - start of periodicity

- Step 3: the onset of voicing is found based on “periodicity”
- A short term autocorrelation computed by multiplying every RTFR frame (for every 0.625 ms frame advance)
- The autocorrelation function: a large value where there is a substantial amount of energy that is periodically repeated with the analysis frame

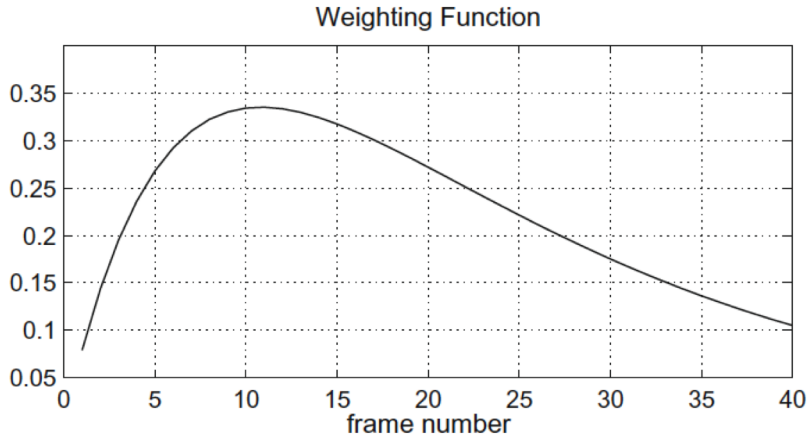


Fig. 3. Weighting function of the periodicity detector.

Experiments - Algorithm performance for phonetic studies

- The absolute difference between the manually and the automatically extracted VOT estimates
  - smaller than 10ms (76.1%), smaller than 20 ms (91.4%), and smaller than 30ms (96.2%)
  - the average deviation: largest for /d/ followed by /k/, /g/, /t/, /p/, and /b/

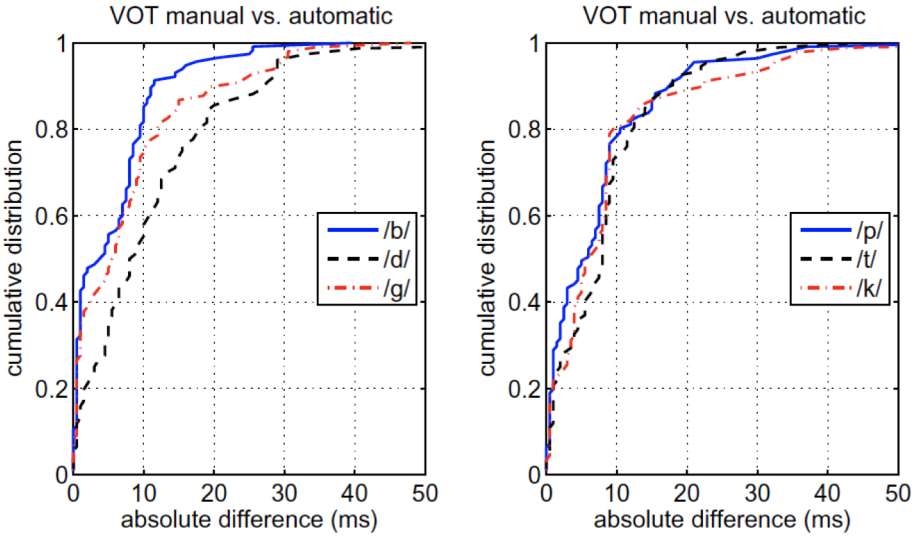


Fig. 5. Absolute difference between the manually and the automatically extracted voice onset time.

Experiments - Algorithm performance for ASR

- The absolute difference between manual and automatic estimates analyzed on the “free” data set
- The absolute difference between the manual and fully automatic VOT estimates
  - smaller than 10ms (72.6%), smaller than 20 ms (87.8%), and smaller than 30ms (93.8%)
- Only 16 (=582-566) out of 582 plosives from the “manual” set could not be found automatically, which is far less than 53 (9.2% of 582)

Experiments - Estimated VOTs

- Such factors as gender and phonetic context considered with respect to the voicing dimension, rather than place of articulation
- VOT: female > male
- Right context

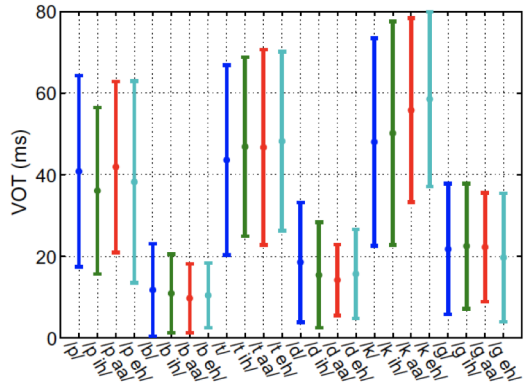


Fig. 6. Mean VOT for plosives /p b t d k g/ by context (context independent, right context /ih/, /aa/, /eh/). The left context is always unconstrained. Error bars indicate +/− one standard deviation. Measured on the "forced" data set.

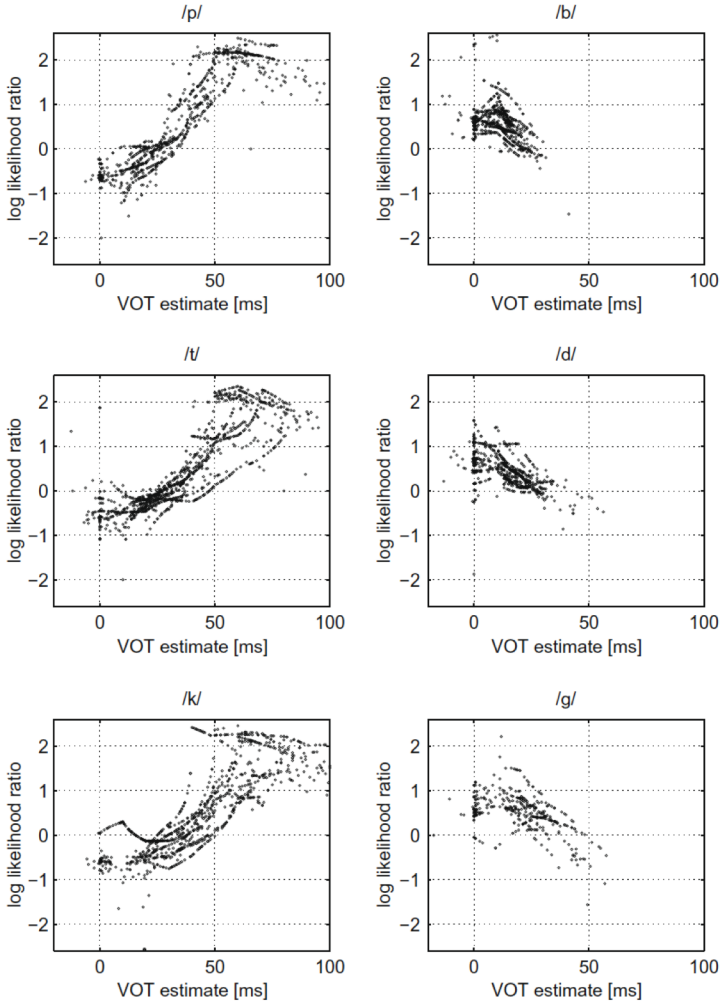
Experiments - VOT as a feature for ASR

- P(V | l, p, r): the probability that the estimated VOT falls in bin V for plosive p
- P(V | l, $\bar{p}$, r): the probability of the plosive with opposite voicing
- The log-likelihood ratio:

$$\log_{10}\left(\frac{P(V|l,p,r)+\varepsilon}{P(V|l,\bar{p},r)+\varepsilon}\right),$$

where

$$P(V|l,p,r)=\frac{N(V,\tilde{l},p,\tilde{r})}{\sum_V N(V,\tilde{l},p,\tilde{r})},$$



Experiments - VOT as a feature for ASR

- In an attempt to improve the phone recognition rate by exploiting the VOT as a feature, phone lattices were generated on the TIMIT test data as described in (Demuynck et al., 2006)
- When dealing with the “test” data set, the left and right phonetic contexts are unique; thus, the set of phone labels of arcs ending (or staring) in the starting (or ending) node of arc A with L (or R) and sum the statistics over all contexts of A allowed by the lattice

$$P(V|\mathscr{L},p,\mathscr{R}) = \frac{\sum_{l\in\mathscr{L}}\sum_{r\in\mathscr{R}}N(V,\tilde{l},p,\tilde{r})}{\sum_{l\in\mathscr{L}}\sum_{r\in\mathscr{R}}\sum_{V}N(V,\tilde{l},p,\tilde{r})}.$$

The corrected acoustic likelihood of a lattice arc A becomes:

$$L(A) + \alpha\log_{10}\left(\frac{P(V|\mathscr{L},p,\mathscr{R})+\varepsilon}{P(V|\mathscr{L},\bar{p},\mathscr{R})+\varepsilon}\right). \quad (1)$$

Experiments - VOT as a feature for ASR

- The single free parameter α was tuned on the “forced” data set, which reduced the phone error rate from 26.70% to 26.53% on the TIMIT test set
- Performance assessment
  - Contributed only very little to error rate improvement (26.70% ⇒ 26.53%)
  - The best obtainable error rate by correcting the voicing of the plosives in the first best path through the phone lattice using the reference transcription ⇒ 25.85%
  - (26.7-26.53)/(26.7-25.85) = 20% of the performance gain achievable using an ideal voicing detector