



**HAL**  
open science

## Automatic Voice Onset Time Estimation from Reassignment Spectra

Veronique Stouten, Hugo Van Hamme

► **To cite this version:**

Veronique Stouten, Hugo Van Hamme. Automatic Voice Onset Time Estimation from Reassignment Spectra. *Speech Communication*, 2009, 51 (12), pp.1194. 10.1016/j.specom.2009.06.003. hal-00575230

**HAL Id: hal-00575230**

**<https://hal.science/hal-00575230>**

Submitted on 10 Mar 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Accepted Manuscript

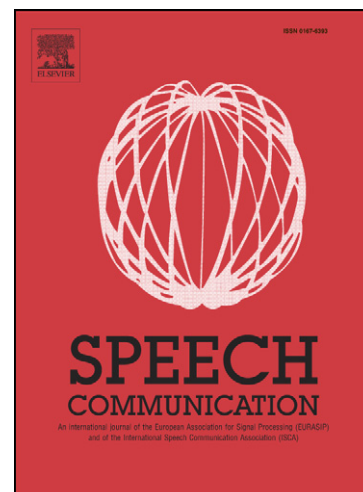
Automatic Voice Onset Time Estimation from Reassignment Spectra

Veronique Stouten, Hugo Van hamme

PII: S0167-6393(09)00089-2  
DOI: [10.1016/j.specom.2009.06.003](https://doi.org/10.1016/j.specom.2009.06.003)  
Reference: SPECOM 1816

To appear in: *Speech Communication*

Received Date: 17 July 2008  
Revised Date: 28 May 2009  
Accepted Date: 1 June 2009



Please cite this article as: Stouten, V., hamme, H.V., Automatic Voice Onset Time Estimation from Reassignment Spectra, *Speech Communication* (2009), doi: [10.1016/j.specom.2009.06.003](https://doi.org/10.1016/j.specom.2009.06.003)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

# Automatic Voice Onset Time Estimation from Reassignment Spectra

Veronique Stouten, Hugo Van hamme \*

*ESAT department, Katholieke Universiteit Leuven  
Kasteelpark Arenberg 10 PO 2441, B-3001 Leuven, Belgium*

---

## Abstract

We describe an algorithm to automatically estimate the voice onset time (VOT) of plosives. The VOT is the time delay between the burst onset and the start of periodicity when it is followed by a voiced sound. Since the VOT is affected by factors like place of articulation and voicing it can be used for inference of these factors. The algorithm uses the reassignment spectrum of the speech signal, a high resolution time-frequency representation which simplifies the detection of the acoustic events in a plosive. The performance of our algorithm is evaluated on a subset of the TIMIT database by comparison with manual VOT measurements. On average, the difference is smaller than 10 ms for 76.1% and smaller than 20 ms for 91.4% of the plosive segments. We also provide analysis statistics of the VOT of /b/, /d/, /g/, /p/, /t/ and /k/ and experimentally verify some sources of variability. Finally, to illustrate possible applications, we integrate the automatic VOT estimates as an additional feature in an HMM-based speech recognition system and show a small but statistically significant improvement in phone recognition rate.

*Key words:* Voice Onset Time, speech attributes, estimation, reassignment spectrum, lattice rescoring.

---

## 1 Introduction

2 State-of-the-art automatic speech recognition (ASR) systems typically use a sliding  
3 window with a length of about 30 ms and a shift of about 10 ms to extract features  
4 such as Mel Frequency Cepstral Coefficients (MFCCs) from the acoustic waveform  
5 of the speech signal. However, plosives also exhibit distinctive acoustic events at a

---

\* Corresponding author. Tel: +32 16 321842, Fax: +32 16 321723.

*Email address:* Hugo.Vanhamme@esat.kuleuven.be (Hugo Van hamme).

6 finer time scale. Typically, the closure interval ends in an abrupt increase in acous-  
7 tic energy across the frequency range. The release interval is measured from this  
8 burst onset to the start of periodicity or to the onset of noise or silence. The duration  
9 of the release interval is then called voice onset time or VOT in case periodicity is  
10 present. These events can be as short as a few milliseconds. Nevertheless, they con-  
11 tain potentially important information on the plosive identity which is lost when  
12 a sliding window of the mentioned size is used. The subsampling caused by the  
13 10 ms frame shift is too slow to accurately represent the timing of the events that  
14 define the release interval and the window length is too large to accurately resolve  
15 the very distinct phases of the plosive. The length of the sliding window and the  
16 frame rate that are used by today's ASR systems are a global compromise on all  
17 phones, involving e.g. effects of the variance of the spectral estimator, the trade-off  
18 between temporal and frequency resolution as dictated by the Heisenberg inequal-  
19 ity, the data rate and the modelling constraints imposed by the subsequent acoustic  
20 modelling techniques such as Hidden Markov Models (HMMs).

21 Recently, there has been considerable interest in supplementing ASR systems with  
22 information that is lost during frame-based front-end processing or that is difficult  
23 to model with popular methods such as HMMs or (hybrid) Multilayer Perceptrons  
24 (Lee et al., 2007). For instance, the phone or state duration distributions implied in  
25 an HMM match poorly with actual distributions measured on speech. In general,  
26 timing at different scales is poorly modeled in traditional ASR systems. Minor ASR  
27 accuracy improvements were found with phone duration models by Seppi et al.  
28 (2007), but the elapsed time between acoustic events at the smallest scale such as  
29 in the current VOT study, or at larger scales such as for prosodic breaks seem to be  
30 difficult to integrate in an ASR system. The work reported in Lee et al. (2007) also  
31 illustrates that the exploitation of speech attributes like the VOT is a substantial  
32 piece of research.

33 The emphasis of this paper is on the automatic measurement of the VOT itself in-  
34 cluding an accuracy analysis. The fact that VOT is not a frame-synchronous feature  
35 but that it is measured at the phone level and that it is only relevant for a subset of  
36 phones makes direct integration in an HMM architecture difficult. However, though  
37 we realize that this is a suboptimal approach, we will illustrate the usefulness of the  
38 VOT feature by rescoreing phone lattices generated by an HMM-based phone recog-  
39 niser. Newer statistical modelling frameworks such as graphical models (Bilmes  
40 and Bartels, 2005) probably offer additional opportunities for more rigorous ap-  
41 proaches to exploit information sources of the type of the VOT. The complexity  
42 of the dependencies on various parameters like gender and phonetic context will  
43 therefore also be described experimentally.

44 Apart from applications in ASR, the current automatic VOT estimator can also be  
45 of interest in speech analysis, phonetics and speech pathology.

46 Acoustic information relevant to the identification of plosive sounds has been stud-

47 ied in the literature (O'Brien, 1993; Whiteside et al., 2004; McCrea and Morris,  
48 2005; Jiang et al., 2006). Plosive consonants are produced by first forming a com-  
49 plete closure in the vocal tract via a constriction at the place of articulation, during  
50 which there is either silence or a low-frequency hum (called voicebar / prevoicing).  
51 The vocal tract is then opened, suddenly releasing the pressure built up behind the  
52 constriction. This opening of the vocal tract's airway is manifested acoustically by  
53 a transient and/or a short-duration noise burst. The duration of the interval between  
54 the release of the plosive and the beginning of voicing in the vowel is called the  
55 voice onset time or VOT. During this interval there is silence and/or noise caused  
56 by the release and/or aspiration noise. The VOT is one of the many acoustic cues  
57 for distinguishing plosives. The acoustic cues relevant to the articulation of a plo-  
58 sive can be related to manner (plosive, nasal, ...), place (bilabial, alveolar, velar,  
59 ...) and voicing (voiced, voiceless). A comprehensive discussion of these cues can  
60 be found in chapter 5 of Borden and Harris (1984) and we limit ourselves to an  
61 enumeration here. The *manner cues* for plosives include the presence of the silent  
62 region in the stop gap (obstruction phase), the rapid formant transitions and partic-  
63 ularly a low locus frequency for the first formant F1, sudden energy change, release  
64 burst and aspiration. The *place cues* for plosives include the burst centre frequency  
65 (i.e. the main spectral peak of the turbulence occurring at the release), the locus  
66 frequency for the second and third formant transitions and the VOT. The *voicing*  
67 *cues* for plosives include the VOT, the presence of aspiration, the presence of an  
68 audible F1 transition, the intensity of the burst and the duration of the preceding  
69 vowel.

70 In this paper, we describe a VOT estimation algorithm using a high resolution sig-  
71 nal analysis method which will better preserve timing information than MFCCs  
72 can. The next section is devoted to this signal representation, the reassigned time-  
73 frequency representation (RTFR). This representation allows well-separated im-  
74 pulses, cosines and chirps to be precisely located in time and in frequency. Because  
75 speech can to some extent be seen as a sum of such signals, we advocate the use  
76 of this representation for our current task. In section 3, the VOT characteristics  
77 are highlighted. A VOT estimation algorithm starts with indentifying segments of  
78 speech that potentially contain a plosive sound. We therefore describe our plosive  
79 data sets in section 4 and move on to section 5 where the actual algorithm that com-  
80 putes the VOT feature from the RTFR is described. Although the VOT has already  
81 been studied extensively, there are not many algorithms described to *automatically*  
82 extract this feature. Related work can be found in Lefebvre and Zwierzynski (1990);  
83 Ramesh and Niyogi (1998); Niyogi and Ramesh (1998); Sonmez et al. (2000);  
84 Kazemzadeh et al. (2006). However, to our knowledge this is the first time that  
85 the RTFR has been used to reliably extract the VOT feature. The performance of  
86 our algorithm is evaluated in section 6.1, while section 6.3 illustrates the modelling  
87 complexity as well as the usefulness of our automatic VOT extraction algorithm  
88 for phonetic studies by measuring some statistics of the VOT feature on the TIMIT  
89 database. Finally, in section 6.4 a rescoreing approach shows a modest improvement

90 in speech recognition accuracy using VOT. Conclusions can be found in section 7.

## 91 2 Spectral reassignment

92 Time-frequency reassignment (Auger and Flandrin, 1995; Plante et al., 1998; Hainsworth  
 93 and Macleod, 2003) offers an interesting solution for analysing transient signals  
 94 such as plosives. The corresponding reassigned time-frequency representation (RTFR)  
 95 has an increased sharpness of localisation of the signal components without sacri-  
 96 ficing the frequency resolution. The RTFR is obtained by moving the spectral den-  
 97 sity value away from the point in the time-frequency plane where it was computed.  
 98 The spectral density is reallocated from the geometric centre of the spectral analy-  
 99 sis kernel function to the centre of gravity of the energy distribution. Though this  
 100 principle can be applied to a multitude of time-frequency representations, here it  
 101 is applied to the short time Fourier transform (STFT). Let  $H(t, \omega)$ ,  $D(t, \omega)$  and  
 102  $T(t, \omega)$  denote the STFT of the signal obtained with the window function  $h(t)$ , the  
 103 derivative of  $h(t)$  and its time-weighted version  $th(t)$  respectively and let  $\Re(X)$  and  
 104  $\Im(X)$  be the real and imaginary parts of  $X$ , then the energy at  $(t, \omega)$  is reassigned  
 105 to:

$$\hat{t} = t - \Re\left(\frac{T(t, \omega)}{H(t, \omega)}\right)$$

$$\hat{\omega} = \omega + \Im\left(\frac{D(t, \omega)}{H(t, \omega)}\right)$$

106 In practical implementations, the time-frequency plane is overlaid with a grid and  
 107 reassigned energy is accumulated per cell.

108 In case the signal is a single cosine, linear chirp or Dirac impulse, the localisation  
 109 in time and frequency is perfect. For instance, for a Dirac impulse  $\delta(t - t_0)$  all  
 110 energy will be reassigned to  $t_0$ . When applied to speech with a sufficiently short  
 111 analysis window, the RTFR clearly shows vertical (i.e. well-localized in time) lines  
 112 for plosive bursts as well as for energy releases by the vocal folds. This property  
 113 will make the construction of detectors for the burst onset of a plosive and for the  
 114 subsequent start of periodicity (if any) fairly easy, as will be shown below. We have  
 115 experimented with the multi-taper version of the RTFR (Xiao and Flandrin, 2007),  
 116 but a single window seemed to provide sufficient detail of the plosives to reliably  
 117 reveal the acoustic events of interest, while it is computationally less demanding.  
 118 Given the impulsive nature of the acoustic events we are trying to characterize, we  
 119 opt for a Hamming window of length 8 ms, shifted by 0.625 ms per analysis frame.  
 120 This corresponds to 128 and 10 samples respectively at a sampling frequency of  
 121 16 kHz which is adopted throughout this paper. Compared to the typical window  
 122 lengths of 20 to 30 ms with a frame advance of 10 ms which are mostly used in

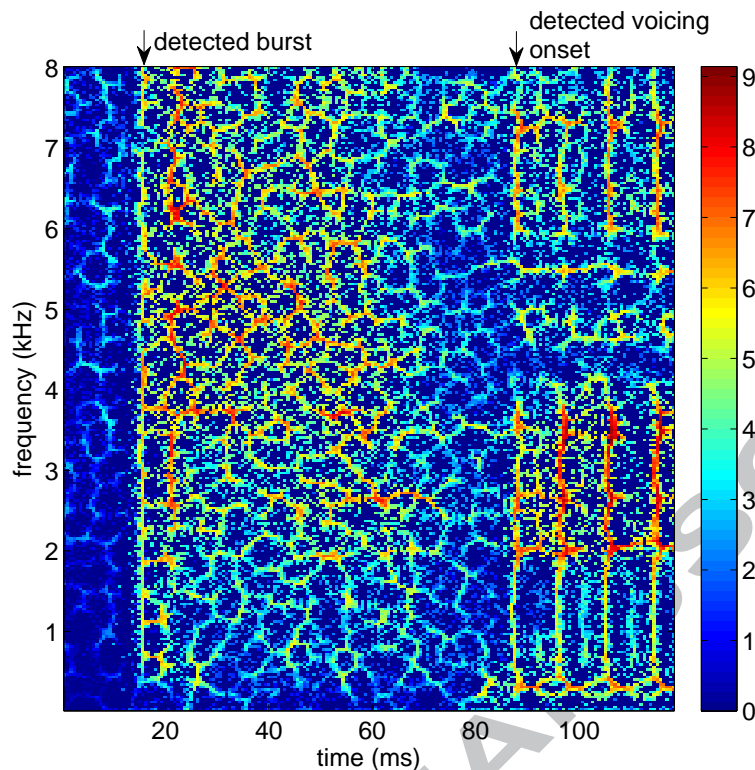


Fig. 1. Reassigned time-frequency representation of a /t/ segment followed by /ih/. Colors encode the logarithm of the energy.

123 speech recognition, our signal analysis offers a higher resolution in time. We used  
 124 256 equally spaced frequency bins for reassignment, a choice which is not critical  
 125 given the wideband nature of the variables upon which the detection of the burst  
 126 and the voicing onset will be based.

127 Figure 1 shows an example of the RTFR for a voiceless plosive (/t/) segment (fol-  
 128 lowed by the vowel /ih/ as in "pit"), taken from the TIMIT database. The burst and  
 129 onset of voicing as detected by the algorithm described in this paper are shown with  
 130 arrows at the top. In this example, the burst of the /t/ is located at 15 ms, while the  
 131 voicing starts at 87 ms, such that the VOT has a value of 62 ms. For comparison,  
 132 we also show the original STFT from which the RTFR is computed in figure 2.  
 133 Clearly, both the alveolar burst and the effects of glottal activity are better localized  
 134 in time in the RTFR.

### 135 3 Properties of the Voice Onset Time

136 On average, the VOT of voiceless plosives is larger than the VOT of voiced plosives,  
 137 and the VOT increases from a bilabial to an alveolar and to a velar stricture. Hence,  
 138 on average we have :



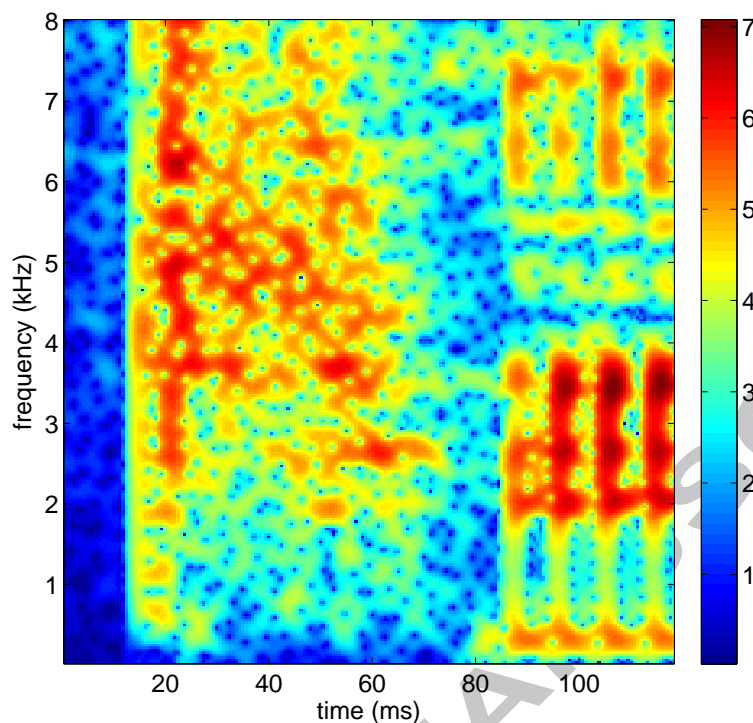


Fig. 2. STFT representation of the /t/ segment from figure 1. Color encode the logarithm of the energy.

$$\text{VOT}(/b d g/) < \text{VOT}(/p t k/)$$

139  $\text{VOT}(/b/) < \text{VOT}(/d/) < \text{VOT}(/g/)$

$$\text{VOT}(/p/) < \text{VOT}(/t/) < \text{VOT}(/k/)$$

140 From the literature, we know that the VOT is influenced by several factors: the  
 141 left and right context of the plosive, the position within the word, the lexical stress,  
 142 speaker gender, speaking rate, the language, fundamental frequency  $F_0$  of the vowel, ... For  
 143 instance, there are notable differences in voicing across languages: Spanish has neg-  
 144 ative VOTs for the voiced plosives, while the VOTs of English are mostly positive.  
 145 Women produce longer VOT values for voiceless stops than men (Whiteside et al.,  
 146 2004). Also, the VOT of children slightly changes with their age. When the plosive  
 147 is followed by the vowel /i/, the mean VOT is larger than when it is followed by  
 148 the vowel /a/ (Whiteside et al., 2004). An increase of the speaking rate causes a  
 149 decrease of the VOT of voiceless plosives. Voiceless stops produced at a high fun-  
 150 damental frequency display shorter VOTs than those at low or mid  $F_0$ 's (McCrea  
 151 and Morris, 2005). In addition, voiceless stops tend to display shorter VOTs and  
 152 voiced stops display increased VOTs during conversational speech and reading,  
 153 compared with isolated words.

154 Because of these effects, VOT distributions tend to overlap. Hence, the relation be-



155 tween the VOT value and plosive identity or even its voicing is not straightforward.  
156 Many studies try to circumvent this overlap by only considering plosives that are  
157 uttered in a constrained way, e.g. single words with a plosive in syllable initial pre-  
158 stressed position. In this way, the variability of the VOT within one class of plosives  
159 becomes smaller. In section 6, it will be shown that statistical models of the VOT  
160 are more precise when they are conditioned on the phonetic context. If these mod-  
161 els are to be used for accuracy gains in ASR as in section 6.4, the context can be  
162 assumed available (although not with 100% accuracy) from a first recognition pass  
163 when evaluating the estimated VOT. By using this knowledge, the overlap of the  
164 distributions can also be reduced to some extent.

## 165 4 Data sets

166 Experiments are conducted on the TIMIT database (Garofolo et al., 1990) since it  
167 contains manually verified phonetic transcriptions. It contains English read speech  
168 at office recording quality, uttered by native adults selected from eight dialect re-  
169 gions in the USA and sampled at a sampling frequency of 16 kHz. Though the  
170 algorithm may also apply to other plosives and affricates, this study focuses on the  
171 six plosives /p/, /t/, /k/, /b/, /d/ and /g/.

172 To study the quality of the VOT estimation algorithm that will be specified (in  
173 section 5), we adopt four data sets that are referred to as "forced", "manual", "free"  
174 and "test". Each of these sets contains a collection of segments of speech in which  
175 we expect to find one of the six plosives. Depending on the data set, the segment  
176 identity as well as its boundaries are generated in different ways as described below.  
177 The number of speech segments for each plosive is given in table 1.

### 178 4.1 The "forced" data set

179 The "forced" data set is relevant for phonetic studies, for automated studies of the  
180 parameters affecting the VOT or for automated pronunciation scoring in (foreign)  
181 language learning. In these settings, speech segments can be found in which one  
182 of the plosives under study is present and our task is to estimate the VOT. The  
183 segment boundaries are obtained from a forced alignment with an HMM-based  
184 speech recogniser using the manually verified phonetic transcriptions available in  
185 the TIMIT database. Hence, we rely on information that is normally not available in  
186 an automatic speech recognition system. All 16134 occurrences of the six plosives  
187 from the 3696 phonetically rich "si" and "sx" training utterances originating from  
188 462 different speakers in the TIMIT database are included in the "forced" data set,  
189 irrespective of the left and right phonetic context.

190 The acoustic models used for segmentation are context independent HMMs with  
191 2 to 4 states per phone trained on an independent data set. In total, there are 141  
192 GMMs sharing 5550 Gaussians with diagonal covariance. The speech features are  
193 mel-scaled log-filterbank outputs that are linearly transformed with a decorrelating  
194 and diagonalizing transform (Demuynck, 2001). Since these features are recalculated  
195 every 10 ms, this is also the segmentation resolution. Voiced plosives and  
196 voiced affricates share a common 2-state HMM for the closure. The voiceless plosives  
197 and affricates also share their closure model. By including separate models  
198 for the phone components of plosives, the HMM will produce separate segments  
199 for the closure and the burst. The segment boundaries that are associated with the  
200 plosive are those of the burst only. The reason for this choice is that the segment  
201 boundaries generated by the HMM will serve as a fallback in case we fail to detect  
202 the burst or the onset of voicing, while the duration of the burst segment can be  
203 seen as a measurement of the VOT.

#### 204 4.2 *The "free" data set*

205 In a fully automatic VOT extraction setting, a forced alignment is not possible due  
206 to the lack of a unique transcription hypothesis. Therefore, in the second data set,  
207 plosive segment candidates are generated by a phonetic automatic speech recogniser  
208 as described in Demuynck et al. (2006) applied to the same utterances used  
209 in the "forced" data set. The HMMs described in section 4.1 are used to find the  
210 best matching phonetic transcription using a phone-level bigram language model  
211 with Witten-Bell smoothing (Witten and Bell, 1991). Any segment automatically  
212 labeled as the burst of one of the six plosives under study was included in the set,  
213 irrespective of the detected phone or phone component on the left and on the right.

#### 214 4.3 *The "manual" data set*

215 The performance of the algorithm will be evaluated by comparing the automatic  
216 VOT estimates with values derived by an expert. To this end, a subset of the plosive  
217 speech segments was selected from the "forced" set as follows. Cycling through  
218 all 16 gender/dialect combinations, we randomly drew a speaker from that gender/  
219 dialect combination and subsequently we randomly drew a recording (sample  
220 file) from that speaker. For any of the six plosives for which we collected less than  
221 130 examples so far, the expert manually estimated the VOT of all occurrences in  
222 the recording by inspection of waveforms and spectrograms centered around the  
223 automatically generated segment boundaries, marking the burst onset time and the  
224 start of voicing and finally storing the time difference. In total 268 different recording  
225 files from the TIMIT database were used. All plosive segments that were not  
226 followed by a voiced sound or for which the manual annotator could not detect a

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	15980	582	4908

227 burst or the start of voicing were removed. There is no constraint on the left pho-  
 228 netic context. Table 1 shows the exact number of examples thus retained in the  
 229 "manual" data set.

#### 230 4.4 The "test" set

231 This set is constructed exactly like the "forced" data set, except that the sentences  
 232 are taken from the TIMIT test set ("extended" set without the "core" set), a total of  
 233 1152 sentences from 144 speakers.

## 234 5 The VOT estimation algorithm

235 The VOT estimation algorithm proceeds in three sequential steps. In the subse-  
 236 quent subsections, each of these steps is described in greater detail. First, candidate  
 237 plosive segments are detected and segment boundaries are generated. Secondly,  
 238 the burst onset is detected by peak picking in the acoustic measure called "burst  
 239 power". Thirdly, the start of voicing is found by peak picking in the acoustic mea-  
 240 sure called "periodicity". The estimated VOT is then the elapsed time between the  
 241 estimated burst onset time and the estimated start of periodicity. Figure 4 illustrates  
 242 the acoustic measures the algorithm relies on as well as the outcome of the peak  
 243 picking criteria (described below) used for detecting both events.

244 The procedure has different possible outcomes. First, the plosive detection may fail  
 245 by generating a false alarm or by missing a plosive. This leads to unrecoverable  
 246 errors in the estimated VOT. Second, the generated segment boundaries may de-  
 247 viate too much from the real start or ending such that a different acoustic event is  
 248 identified as the burst or voicing onset. For instance, if the proposed segment start

249 is too early, an event belonging to the previous phone may be identified as the burst  
250 onset. If the proposed segment start is too late, the burst onset may not be detected.  
251 In the latter case, the missed event will be related to the segment boundaries pro-  
252 posed by the detector (see below), but given the erroneous segment boundary, the  
253 VOT error will be important. Third, either burst or voicing may not be revealed  
254 by their acoustic measure, in which case fallback estimates of their time of occur-  
255 rence are derived from the segment boundaries proposed by the plosive detector.  
256 In this case, the VOT errors critically depend on the quality of the generated seg-  
257 ment boundaries. Fourth, the segment may be correctly identified as a plosive with  
258 successful timings of the burst and voicing onset, leading to small VOT estimation  
259 error related to the time-frequency representation.

### 260 5.1 Detection of plosive segments

261 The first step in the algorithm consists of finding segments in the speech signal that  
262 could contain a plosive. Such segments could be found using dedicated detectors,  
263 as is shown in the research on automatic extraction of phonological features. In  
264 King and Taylor (2000) and Stouten and Martens (2006), detectors are described  
265 that exhibit sufficient accuracy to generate candidate plosive segments. The method  
266 used for generating plosive segment candidates is important to the performance of  
267 the algorithm. In the introduction of section 5, four categories of outcomes were  
268 defined. For the first outcome, one needs to optimize the trade-off between false  
269 alarms and missed detections. For the second and third outcome, the proposed seg-  
270 ment boundaries need to be as accurate as possible.

271 In the current work, we have opted for a HMM-based automatic speech recogniser  
272 to generate plosive segment candidates, as explained in section 4. Depending on the  
273 application of the VOT estimate, it may or may not be reasonable to assume that  
274 a phonetic transcription of the speech around the plosive is available. We therefore  
275 defined the "forced" and "free" data sets in which plosive segments are generated  
276 with or without phonetic knowledge of the test utterance. In both sets, the algorithm  
277 will start looking for the burst 2.5 ms or 4 frames prior to the burst segment start  
278 found by the recogniser. Starting earlier would increase the risk of misdetecting  
279 energy bursts from the previous phone as belonging to the plosive. Starting later  
280 would increase the risk of missing the burst. The end of the segment is extended  
281 by 10 ms or 16 frames to the future. Extension of the segment end to the right  
282 just means more pitch cycles will be included and is harmless to the algorithm.  
283 The value of 10 ms is a compromise such that at least one glottal closure will be  
284 seen in most cases, while avoiding unreasonably high VOT estimates in case some  
285 initial glottal vibration cycles are not detected. Notice that even if the successor  
286 segment was manually or especially automatically labeled as a vowel, this does not  
287 *guarantee* that glottal activity will be detected.

288 In the discussion below, we will refer to *extended* segments to refer to the plosive  
 289 segment starting 2.5 ms before and ending 10 ms after the segment determined by  
 290 the speech recogniser.

## 291 5.2 *Burst onset detection*

292 Figure 1 shows that the onset of the release phase gives rise to a sudden increase of  
 293 the amplitude over the whole frequency range.

294 To limit the influence of the high-amplitude pitch pulses which also have a strong  
 295 low-frequency component, only the frequency range 3.2-8 kHz is retained for burst  
 296 detection. The corresponding frequency bins in the RTFR power are summed to  
 297 form the "burst power"  $p(n)$  estimate for frame  $n$ . Then, the first local maximum  
 298 in  $p(n)$  that is sufficiently strong and ramps up sufficiently sharply is identified as  
 299 the burst onset. The condition is asymmetric because  $p(n)$  can stay high during  
 300 the release interval after the burst. In formulae, frame  $n$  is retained as a possible  
 301 burst location if it satisfies all of the following conditions  $p(n) > p(n - j)$ , for  
 302  $j = -1, 1$  and 2 (local maximum),  $p(n) - p(n - i) > p_m(n)$  for  $i = 2 \dots 5$   
 303 (sufficiently sharp and strong peak), where  $p_m(n)$  is a measure that relates to the  
 304 average signal energy so the criteria are invariant to scaling of the signal. In our  
 305 experiments,  $p_m(n)$  is taken to be the mean of  $p(n)$  over 150 plosive frames.

306 If the automatic algorithm does not find a local maximum, the start of the (unex-  
 307 tended) segment is marked as the burst onset. This may happen because the burst  
 308 is simply missing (by construction, this will not happen in the "manual" data set)  
 309 or because it is too weak. The resulting estimate is less accurate: measured over  
 310 all plosives of the "manual" data set, the square root of the mean square estimation  
 311 error is 12.6 ms if a burst was detected, while it increases to 22.6 ms if a burst could  
 312 not be detected.

## 313 5.3 *Start of periodicity*

314 As can be seen from the RTFR in figure 1, the periodicity of the signal gives rise to  
 315 vertical lines of high amplitude with valleys in between. The distance between these  
 316 lines is determined by the pitch period. This periodic structure is mainly present in  
 317 the lower part of the frequency range.

318 To obtain a robust estimate of the start of voicing, only the frequency range 0-4 kHz  
 319 is retained. At a sampling frequency of 16 kHz as used in this work, this comes  
 320 down to keeping only the lower half of the RTFR. Then, a short term autocorrelation  
 321 is computed by multiplying every RTFR frame (for every 0.625 ms frame advance)  
 322 with a weighted version of the frames at lags 1 to 40 and summing these values

323 over the lag index and over the retained frequency bins. The weighting function  
 324 (figure 3) is given by the difference of two decaying exponential functions and has  
 325 a large value in the adult pitch period range of 5 to 20 frames, corresponding to  
 326 a pitch period between 3.1 ms and 12.5 ms or a pitch frequency of 320 Hz down  
 327 to 80 Hz. An asymmetric weighting function is chosen because we want to extract  
 328 the *start* of periodicity. The result is normalised with the total energy in the frames  
 under the autocorrelation window over the whole frequency range (0-8 kHz).

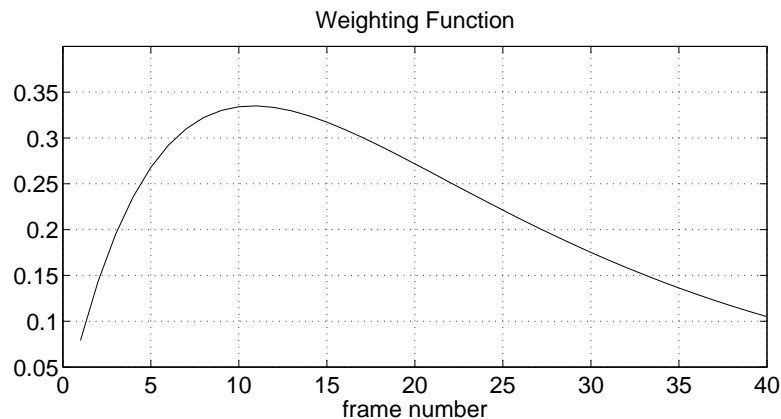


Fig. 3. *Weighting function of the periodicity detector.*

329

330 The autocorrelation function obtained in this way will exhibit a large value at times  
 331 where there is a substantial amount of energy that is periodically repeated within  
 332 the analysis frame, i.e. at the time instants for which a pitch pulse is present in  
 333 the RTFR. To be marked as a local maximum, the following conditions have to be  
 334 met: its value has to be larger than the value of its direct neighbours, and it has to  
 335 exceed the value of its neighbours at distances of  $\pm 2$ ,  $\pm 3$  and  $\pm 4$  frames with  
 336 an increasing threshold to assure that the selected peaks are at least 5 frames (or  
 337 the minimum pitch period) on either side from their neighbours and at least 0.03 in  
 338 height, a value which was determined from visual inspection on the "forced" data  
 339 set (excluding the "manual" set).

340 With this scheme, some of the bursts will also be marked as pitch pulses. Moreover,  
 341 a velar stricture can have multiple bursts that should not be confused with pitch  
 342 pulses. To avoid selecting the burst as the start of voicing, an additional constraint  
 343 is imposed. A local maximum has to be within the maximal pitch period (20 frames  
 344 or 12.5ms) from the *next* local maximum (or from the end of the extended segment).  
 345 For low-pitched voices, the wrong starting point of voicing can still be selected if  
 346 some pitch pulses are not detected. However, the risk of selecting the burst onset is  
 347 strongly reduced, especially if multiple bursts are present.

348 If the algorithm cannot detect voicing within the extended segment, the end of  
 349 the unextended segment is marked as the start of voicing, i.e. we fall back to the  
 350 HMM's decision of the start of the next phone. This is a reasonable choice for  
 351 English, where VOTs are mostly positive, but for other languages, voicing may al-



352 ready start in the closure interval. On the "manual" data set, we measure a square  
 353 root of the mean square error of 12.2 ms if voicing was detected, while it increases  
 354 to 17.8 ms if voicing could not be detected within the extended segment. Not sur-  
 355 prisingly, the HMM does a better job at detecting the start of the next vowel than it  
 356 does at detecting the burst.

#### 357 5.4 Discussion

358 The proposed peak picking algorithms are surely not the only possible approaches  
 359 to detecting the burst and voice onset events in RTFRs. The advantage of the RTFR  
 360 is that the peaks are clear and sharp, which motivates the high time resolution of  
 361 0.625 ms used in our proposed algorithm. Often, both the burst and the glottal  
 362 closures can be located to a single frame. Decreasing the frame rate might make  
 363 the algorithm computationally more efficient, but would make the peak picking  
 364 more error prone. In any case, even at pitch periods down to about 3 ms, sampling  
 365 needs to be fast enough to resolve the pitch peaks. Similarly, the burst onset may  
 366 exhibit multiple clicks which should not be merged into a single broad peak of  $p(n)$   
 367 if the same peak detection criteria are maintained.

## 368 6 Experiments

### 369 6.1 Algorithm performance for phonetic studies

370 The VOT was estimated for the complete "forced" data set by means of the auto-  
 371 matic algorithm of section 5. Since the "manual" data set is a subset of the "forced"  
 372 set, it is possible to compare the manual and automatic VOT estimates on this sub-  
 373 set. Figure 5 shows the cumulative distribution of the absolute difference between  
 374 the manually and the automatically extracted VOT estimates. On average, the dif-  
 375 ference is smaller than 10 ms in 76.1% of the plosive segments, smaller than 20 ms  
 376 for 91.4% of the plosive segments, and smaller than 30 ms for 96.2% of the plosive  
 377 segments. The average deviation from the manually assigned VOT is the largest for  
 378 /d/ and decreases from /d/ to /k/, /g/, /t/, /p/ and /b/.

379 Table 2 gives an indication of the bias of the algorithm. For each plosive, it con-  
 380 tains the average of the manually and of the automatically extracted VOTs on the  
 381 "manual" data set. The resulting bias is calculated as the difference of both means  
 382 and the uncertainty on this estimate is given as its standard deviation assuming in-  
 383 dependent bias measurements. There is an overall bias of 2.9 ms, which is even  
 384 statistically detectable on most individual plosives. To show that the bias is mainly  
 385 due to the fallback in case either burst or voicing onset cannot be detected automat-



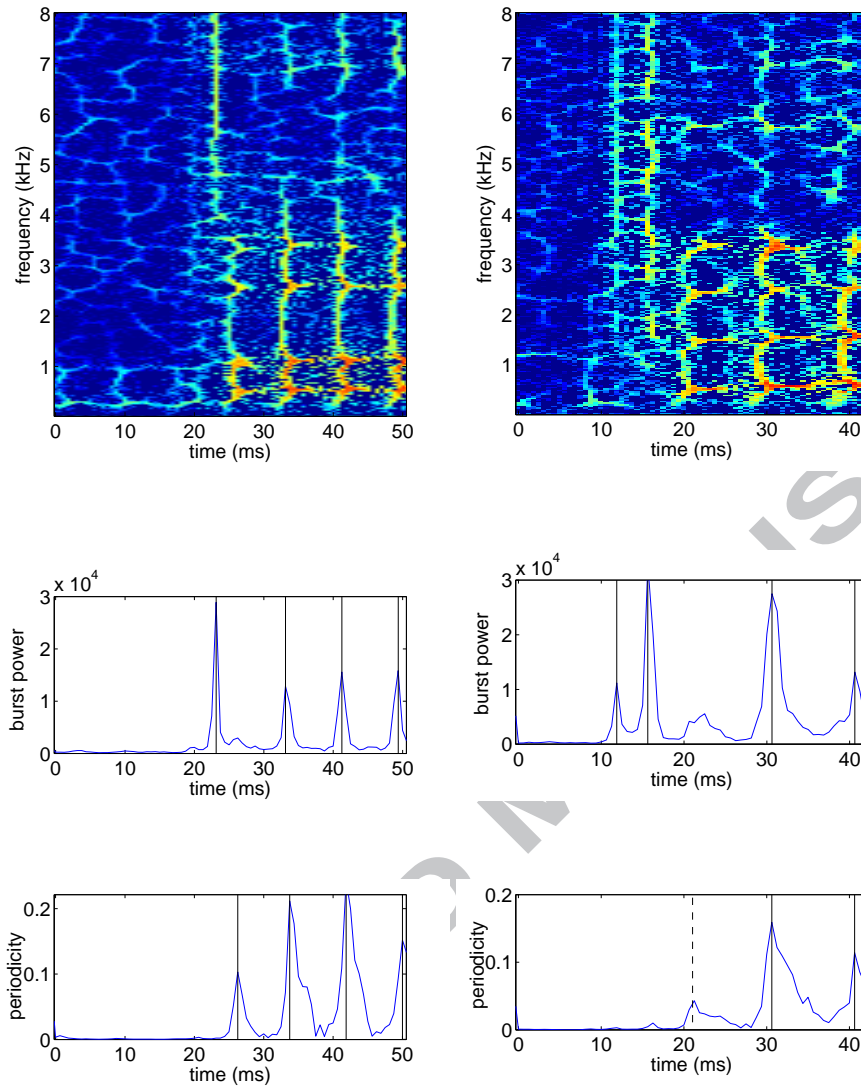


Fig. 4. *Left: illustration of the peak picking on a /b/ segment with a right context /aa/ (from "flat bottom") taken from the "free" data set. From top to bottom: RTFR, burst detection and periodicity detection. The peaks that satisfy the selection criteria are marked with vertical solid lines. Right: /b/ segment (from the word "thereby") with erroneous detection of the start of voicing. The missed start of periodicity is marked with a dashed line.*

386 ically, the right side of the table gives the same statistics measured only on those  
 387 segments from the "manual" data set for which the algorithm was able to detect  
 388 both events. The overall bias is now down to 0.9 ms and mostly realized on /d/. A  
 389 further analysis would need to question the human annotation as well as the peak  
 390 selection criteria. Phenomena as illustrated in the right pane of figure 4 are likely  
 391 to play a role here.

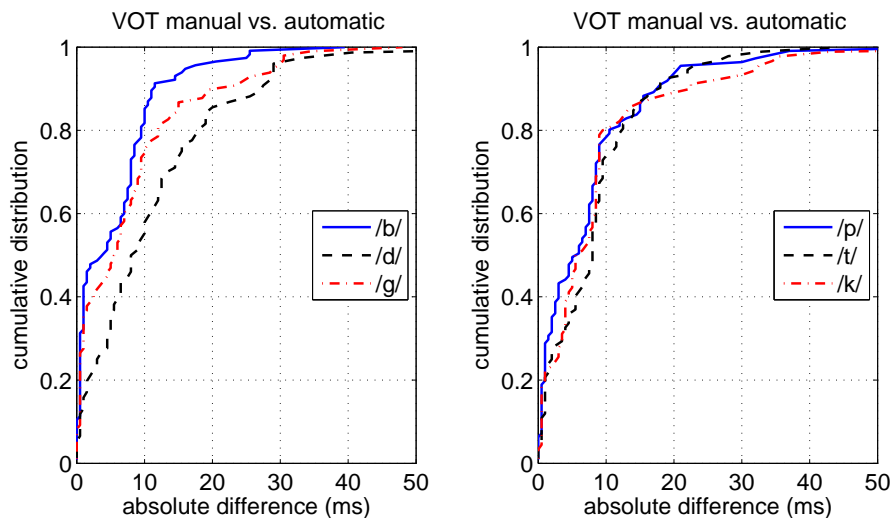


Fig. 5. Absolute difference between the manually and the automatically extracted voice onset time.

Table 2

Comparison between the average manually and automatically extracted VOT for each plosive. Left: all plosive segments of the "manual" data set. Right: only the plosive segments for which both burst and voicing onset could be detected automatically.

	VOT (ms)				VOT (ms)			
	all segments				without fallback			
	manual	autom	bias	stdev	manual	autom	bias	stdev
/b/	7.7	9.8	2.1	0.9	7.9	8.8	0.9	0.8
/d/	8.5	16.1	7.7	1.9	8.2	13.5	5.2	2.0
/g/	21.8	22.7	0.9	1.1	21.7	21.7	0.0	1.1
/p/	39.4	44.1	4.6	1.1	38.5	40.4	1.9	1.2
/t/	50.9	51.4	0.6	1.2	50.2	48.9	-1.3	1.3
/k/	54.3	56.4	2.1	1.7	56.2	55.2	-1.1	2.0
avg	30.3	33.1	2.9	0.5	28.8	29.7	0.9	0.5

## 392 6.2 Algorithm performance for automatic speech recognition

393 While the above accuracy analysis is relevant for e.g. phonetic studies, where seg-  
 394 ment boundaries can be generated based on a manually produced phonetic tran-  
 395 scription, its validity can be questioned in a fully automatic setting, where the goal  
 396 of VOT estimation could be to improve speech recognition accuracy on plosives.  
 397 Therefore, in the second study, the absolute difference between manual and auto-  
 398 matic estimates is analysed on the "free" data set. However, an automatic phone  
 399 recogniser can mislabel plosive segments, insert or omit them, or generate different

Table 3

*VOT estimate [ms] for each plosive class, averaged over all contexts in the "forced" data set. Mean value for all speakers, only male or only female speakers. Columns 5-7 indicate the corresponding number of segments.*

	VOT [ms]			# segments		
	m + f	m	f	m + f	m	f
/b/	11.8	11.3	13.0	2181	1522	659
/d/	18.6	17.7	20.5	2432	1681	751
/g/	21.8	20.7	24.0	1191	800	391
/p/	40.8	39.0	45.0	2588	1798	790
/t/	43.6	41.8	48.1	3948	2791	1157
/k/	48.0	47.1	50.3	3794	2686	1108

400 segment boundaries. We related the plosive segments from the "free" data set with  
 401 one from the "forced" data sets by selecting the "forced" plosive segment with the  
 402 largest overlap in time. For 9.2% of the segments, there was no overlap. Only 0.04%  
 403 of "free" segments overlapped with more than one "forced" segment, in which case  
 404 we took the "forced" plosive with the largest overlap in time. Notice that it may  
 405 well be that the phone identity (among the set of six considered) is different in both  
 406 sets, corresponding to the mislabelings by the recogniser that we are trying to cor-  
 407 rect. In this analysis, the manual phonemic labelings provided the TIMIT database  
 408 are assumed to be correct.

409 With this procedure, 566 plosive segments from the "free" set could be linked  
 410 with a segment from the "manual" set, which allows the cumulative distribution  
 411 of the absolute difference between manual and fully automatic VOT estimates to  
 412 be recomputed. The percentiles for 10 ms, 20 ms and 30 ms deviation now be-  
 413 come 72.6%, 87.8% and 93.8% respectively (instead of 76.1%, 91.4% and 96.2%).  
 414 Hence, the main source of estimation error is not caused by the automatic gener-  
 415 ation of segment boundaries. Also notice that only 16 (= 582 - 566) out of 582  
 416 plosive segments from the "manual" set could not be found automatically, which is  
 417 far less than 53 (9.2 % of 582). Hence, the HMM-based plosive detector performs  
 418 a lot better on plosives for which the human annotator found a burst that are also  
 419 followed by a voiced sound.

### 420 6.3 Estimated VOTs

421 With this automatic algorithm, we can investigate to which extent factors such as  
 422 gender and phonetic context could be taken into account in statistical models. In  
 423 this study, we focus on the voicing dimension, rather than place of articulation.

424 First, we measure the effect of gender. The second column of table 3 shows the VOT  
 425 obtained on the "forced" data set for each of the plosives, averaged over all speakers  
 426 and all contexts. These values confirm the inequalities of section 3. Columns 3 and  
 427 4 contain the VOT values averaged over all contexts but including only the male,  
 428 or only the female speakers, respectively. On our database, the VOTs of plosives  
 429 uttered by women are on average 12% longer than that of men. For /p t k/, this  
 430 is in line with Whiteside et al. (2004), but the latter article did not mention the  
 431 same effect for /b d g/. Notice that the gender-independent averages differ from  
 432 those of table 2 because the phonetic context of the plosives differs, as explained in  
 433 section 4.

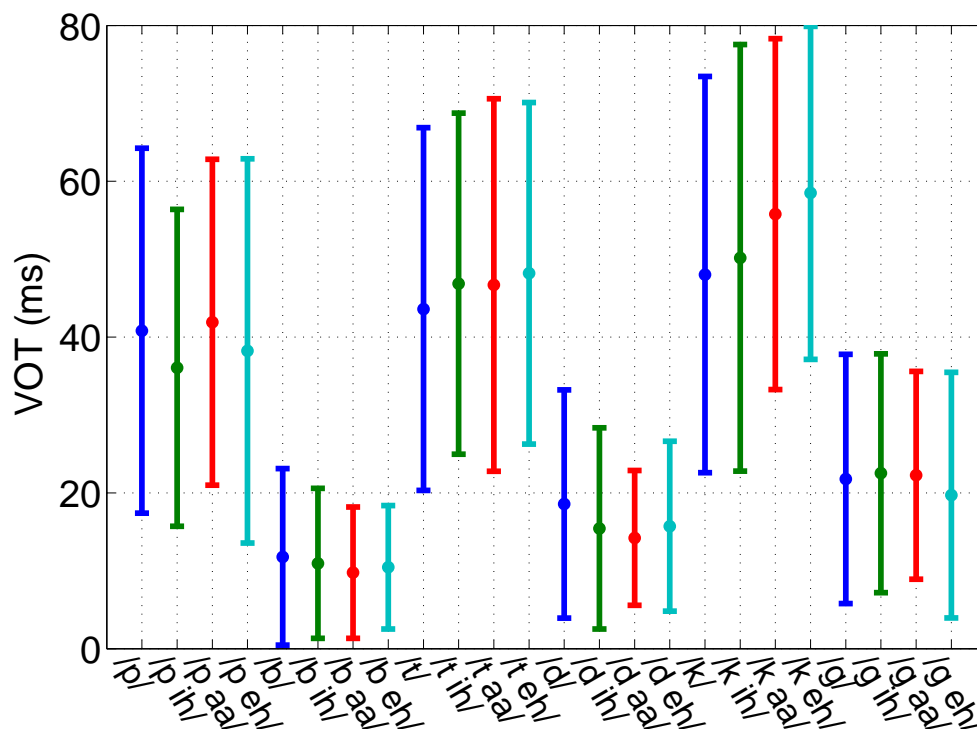


Fig. 6. Mean VOT for plosives /p b t d k g/ by context (context independent, right context /ih/, /aa/, /eh/). The left context is always unconstrained. Error bars indicate +/- one standard deviation. Measured on the "forced" data set.

434 The effect of the right context can be found in figure 6, which presents the VOT  
 435 means together with the standard deviations without any right context imposed or  
 436 when it is followed by a vowel /ih/ (as in "bit"), /aa/ (as in "box") or /eh/ (as in  
 437 "bet"). There is no constraint on the left context. In total, there are between 68  
 438 and 253 examples of each right-context dependent plosive in the database when  
 439 pooling over all speakers. If the phonetic context is constrained, the overlap of the  
 440 VOT distributions usually decreases. For instance, the error bars of /k eh/ and /g eh/  
 441 do not overlap, while the error bars for the context independent /k/ and /g/ do. The  
 442 same can be said about /p aa/ and /b aa/ versus /p/ and /b/. The longer average VOT  
 443 for right context /ih/ than for context /aa/ is only observed for plosives /b d g t/.

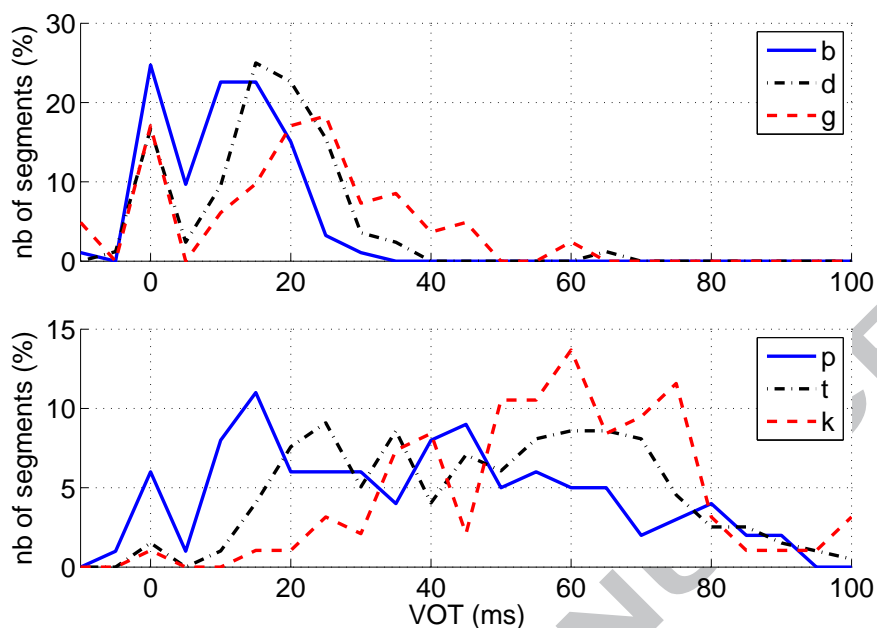


Fig. 7. Normalised histogram of VOT estimates on the "forced" data set for plosives /b d g/ and /p t k/ followed by vowel /eh/, without constraint on the left context.

444 Figure 7 shows histograms of the context dependent VOTs of plosives followed by  
 445 the vowel /eh/, constructed on the "forced" data set. From this figure, the overlap  
 446 of the distributions is clearly apparent. This overlap is even larger for the context  
 447 independent histograms. This illustrates that the relation between the VOT value  
 448 and the voicing cue of the plosive is not straightforward.

#### 449 6.4 VOT as a feature for automatic speech recognition

450 Histograms like the one of figure 7 can be used in a likelihood ratio test to discrim-  
 451 inate, for instance, along the voicing dimension. To this end, context dependent but  
 452 gender independent histograms are built with 23 uniformly spaced bins 5 ms apart  
 453 between -10 ms and +100 ms using the "forced" data set. Let  $N(V, l, p, r)$  be the  
 454 number of times the estimated VOT falls in bin  $V$  for plosive  $p$  with left context  
 455  $l$  and right context  $r$ . Overall, 1298 different phone/plosive/phone combinations  
 456 are observed. Many of these histograms have little data, so a multi-stage backoff  
 457 scheme is applied to histograms having less than 40 counts, i.e. if

$$\sum_V N(V, l, p, r) < 40$$

458 First the left context is generalised to one of 12 broad phonetic classes, then the  
 459 right context is generalized, then the left context is disregarded and finally the right

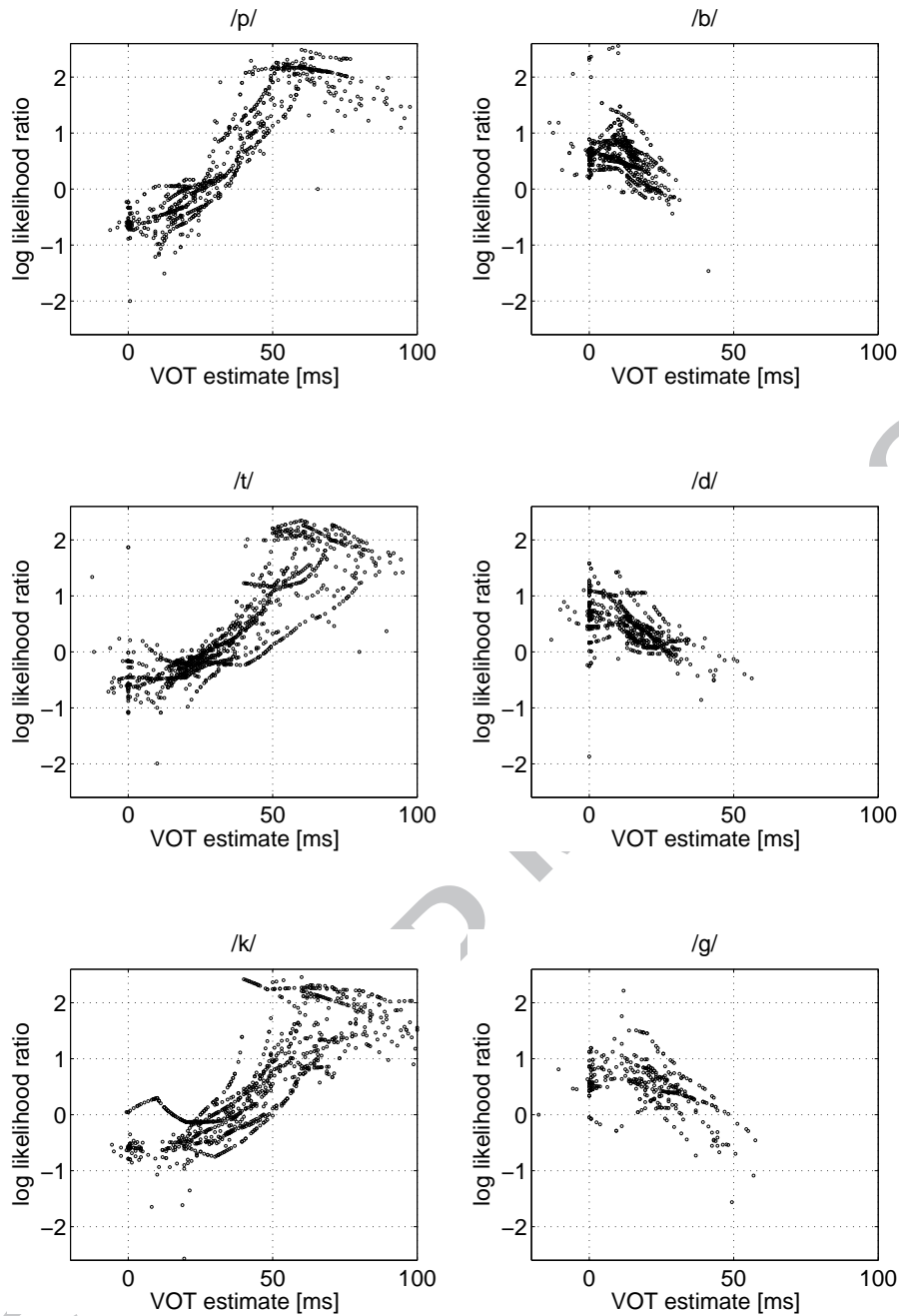


Fig. 8. *Logarithm of the likelihood ratio versus the automatically calculated VOT value, measured on the "test" data set.*

460 context is disregarded. The backoff steps are terminated as soon as at least 40 counts  
 461 are observed in the histogram with the generalized context. We will call the thus  
 462 obtained generalized left and right context  $\tilde{l}$  and  $\tilde{r}$  respectively.

463 Figure 8 shows the logarithm (to base 10) of the likelihood ratio versus the esti-  
 464 mated VOT value for the "test" data set. This set contains data that was not used dur-

465 ing the construction of the histograms, while the ground truth about plosive identity  
 466 and its context is known from the manual labeling provided in the TIMIT database.  
 467 So let  $P(V|l, p, r)$  be the probability that the estimated VOT falls in bin  $V$  for plo-  
 468 sive  $p$  as measured on its histogram, and let  $P(V|l, \bar{p}, r)$  be the probability read  
 469 from the histogram for the plosive with opposite voicing. The log-likelihood ratio  
 470 is then

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

471 where

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

472 and  $\varepsilon$  is a small constant to cope with zero probability estimates and was set to  
 473  $10^{-3}$  in our experiments. The left panes show the log-likelihood ratio on the voice-  
 474 less data and assuming the voiceless sound ( $p$  is /p/, /t/ or /k/ and  $\bar{p}$  is /b/, /d/ or /g/  
 475 respectively), while the right panes show the log of the reciprocal on the voiced  
 476 data (i.e. assuming  $p$  is a voiced sound). Figure 8 illustrates that large (small)  
 477 VOTs for voiceless (voiced) sounds indeed lead to positive log-likelihood ratios,  
 478 but that negative log-ratios can occur. That the choice of  $\varepsilon$  is not a critical one is  
 479 also apparent from these scatter plots. Its side-effect is to limit extreme values of  
 480 the log-likelihood ratio, an effect that is mostly observed on the positive side.

481 In an attempt to improve the phone recognition rate by exploiting the VOT as a  
 482 feature, phone lattices were generated on the TIMIT test data as described in De-  
 483 muynck et al. (2006). These are the same sentences as used in the "test" data set,  
 484 but now the lattice will include more plosive candidates. The best path through the  
 485 lattice will generate the phone segmentation of the "test" data set. In formula 1, the  
 486 likelihood  $L(A)$  of each plosive arc  $A$  in the lattice is then linearly combined with  
 487 the log-likelihood ratio of it being correct versus its variant with opposite voicing  
 488 being correct. There is, however, a difference with the above. When dealing with  
 489 the "test" data set, the left and right phonetic contexts are unique. In a lattice, mul-  
 490 tiple arcs may arrive in the starting node of  $A$  and multiple arcs may leave from its  
 491 ending node, so the left and right phonetic context are not unique. We denote the  
 492 set of phone labels of arcs ending (or starting) in the starting (or ending) node of  
 493 arc  $A$  with  $\mathcal{L}$  (or  $\mathcal{R}$ ) and sum the statistics over all contexts of  $A$  allowed by the  
 494 lattice:

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



495 The corrected acoustic likelihood of a lattice arc  $A$  becomes:

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

496 Linear combination of log-likelihoods of different information sources was exam-  
 497 ined in Beyerlein (1998). The single free parameter  $\alpha$  we introduced was tuned on  
 498 the "forced" data set, which is independent of the "test" data set. This procedure re-  
 499 duced the phone error rate from 26.70% to 26.53% on the TIMIT test set. Hence, we  
 500 observe that the VOT feature has contributed only very little to error rate improve-  
 501 ment. This is not surprising, since we observe in figure 8 that the log-likelihood  
 502 ratio can become negative for valid utterances of the plosive. On the other hand we  
 503 have to realize that we attempt to correct only the plosive hypotheses generated by  
 504 the HMM system, and this only along the voicing dimension. We can find the best  
 505 obtainable error rate by correcting the voicing of the plosives in the first best path  
 506 through the phone lattice using the reference transcription. This yields an error rate  
 507 floor of 25.85%. Hence, we have obtained  $(26.7 - 26.53)/(26.7 - 25.85) = 20\%$   
 508 of the performance gain that would be achievable using an ideal voicing detec-  
 509 tor. In absolute numbers, the VOT-based likelihood ratio test corrected 80 out of  
 510 1853 plosive errors and hence the improvement is statistically significant. The gain  
 511 shows that the VOT estimate does contain information that the HMM is not able  
 512 to exploit. Apart from the overlap in the distributions of the VOT, the performance  
 513 in this particular implementation is also limited by the pruning in the phone lattice.  
 514 Each plosive hypothesis (arc) is rescored, but this can only lead to a change in de-  
 515 cision if the hypothesis with opposite voicing is also in the lattice (and receives a  
 516 better combined score). Hence, if the alternate, correct hypothesis was not included  
 517 in the lattice because of pruning, it cannot be recovered, even with an ideal voicing  
 518 detector. Further performance improvements might also be obtained by combining  
 519 the HMM and VOT likelihoods in a non linear way.

## 520 7 Conclusions

521 We have described an algorithm to *automatically* extract the voice onset time. It op-  
 522 erates on the reassigned time-frequency representation of the signal, which has an  
 523 improved localisation of the relevant acoustic events. The algorithm performance  
 524 was characterised for English plosives on the TIMIT database. The accuracy seems  
 525 sufficient to reconstruct and extend some of the findings of the phonetics literature  
 526 about the factors affecting VOT. Using a rescoreing approach, it was shown that the  
 527 automatic VOT estimate does provide some additional information about the phone  
 528 identity which is not exploited in state-of-the-art HMM-based ASR systems.

529 **8 Acknowledgement**

530 This research was funded by the IWT - SBO project 'SPACE' (project no.040102)  
531 and by the European Commission under contract FP6-034362 (ACORNS).

532 **References**

- 533 Auger, F., Flandrin, P., 1995. Improving the readability of time-frequency and time-  
534 scale representations by the reassignment method. *IEEE Trans. on SP* 43 (5),  
535 1068–1089.
- 536 Beyerlein, P., May 1998. Discriminative model combination. In: *Proc. ICASSP*.  
537 Seattle, WA, U.S.A., pp. 481–484.
- 538 Bilmes, J., Bartels, C., 2005. Graphical model architectures for speech recognition.  
539 *IEEE SP Mag.* 22 (5), 89–100.
- 540 Borden, G. J., Harris, K. S., 1984. *Speech Science Primer: Physiology, Acoustics,*  
541 *and Perception of Speech*, 2nd Edition. Williams & Wilkins, Baltimore, U.S.A.
- 542 Demuynck, K., Feb. 2001. Extracting, modelling and combining information in  
543 speech recognition. Ph.D. thesis, K.U.Leuven, Belgium.
- 544 Demuynck, K., Van Compernelle, D., Van hamme, H., Sep. 2006. Robust phone  
545 lattice decoding. In: *Proc. ICSLP*. Pittsburgh, U.S.A., pp. 1622–1625.
- 546 Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V.,  
547 Oct. 1990. The DARPA TIMIT acoustic-phonetic continuous speech corpus. In:  
548 *Speech Disc 1-1.1*, NTIS Order No. PB91-5050651996.
- 549 Hainsworth, S., Macleod, M., 2003. Time-frequency reassignment: a review and  
550 analysis. Tech. Rep. CUED/FINFENG/TR.459, Cambridge University Engineer-  
551 ing Department.
- 552 Jiang, J., Chen, M., Alwan, A., 2006. On the perception of voicing in syllable-initial  
553 plosives in noise. *J. of ASA* 119 (2), 1092–1105.
- 554 Kazemzadeh, A., Tepperman, J., Silva, J., You, H., Lee, S., Alwan, A., Narayanan,  
555 S., Sep. 2006. Automatic detection of voice onset time contrasts for use in pro-  
556 nunciation assessment. In: *Proc. ICSLP*. Pittsburgh, PA, U.S.A.
- 557 King, S., Taylor, P., 2000. Detection of phonological features in continuous speech  
558 using neural networks. *Comp. Speech and Lang.* (14), 333–353.
- 559 Lee, C.-H., Clements, M. A., Dusan, S., Fosler-Lussier, E., Johnson, K., Juang,  
560 B.-H., Rabiner, L., Sep. 2007. An overview on automatic speech attribute tran-  
561 scription (asat). In: *Proc. INTERSPEECH*. Antwerp, Belgium, pp. 1825–1829.
- 562 Lefebvre, C., Zwierzynski, D., Nov. 1990. The use of discriminant neural networks  
563 in the integration of acoustic cues for voicing into a continuous-word recognition  
564 system. In: *Proc. ICSLP*. Kobe, Japan, pp. 1073–1076.
- 565 McCrea, C., Morris, R., 2005. The effects of fundamental frequency level on voice  
566 onset time in normal adult male speakers. *J. of Speech, Lang. and Hearing Res.*  
567 48, 1013–1024.

- 568 Niyogi, P., Ramesh, P., May 1998. Incorporating voice onset time to improve letter  
569 recognition accuracies. In: Proc. ICASSP. Seattle, WA, U.S.A., pp. 13–16.
- 570 O’Brien, S., 1993. Spectral features of plosives in connected-speech signals. *Int. J.*  
571 *Man-Mach. Studies* 38, 97–127.
- 572 Plante, F., Meyer, G., Ainsworth, W., 1998. Improvement of speech spectrogram  
573 accuracy by the method of reassignment. *IEEE Trans. on SAP* 6 (3), 282–286.
- 574 Ramesh, P., Niyogi, P., Nov. 1998. The voicing feature for stop consonants: Acous-  
575 tic phonetic analyses and automatic speech recognition experiments. In: Proc.  
576 ICSLP. Sydney, Australia.
- 577 Seppi, D., Falavigna, D., Stemmer, G., R., G., Sep. 2007. Word duration modeling  
578 for word graph rescoring in lvcsr. In: Proc. INTERSPEECH. Antwerp, Belgium,  
579 pp. 1805–1808.
- 580 Sonmez, K., Plauche, M., Shriberg, E., Franco, H., Oct. 2000. Consonant discrim-  
581 ination in elicited and spontaneous speech: a case for signal-adaptive front ends  
582 in ASR. In: Proc. ICSLP. Beijing, China.
- 583 Stouten, F., Martens, J.-P., Sep. 2006. Speech recognition with phonological fea-  
584 tures: Some issues to attend. In: Proc. ICSLP. Pittsburgh, PA, U.S.A., pp. 357–  
585 360.
- 586 Whiteside, S., Henry, L., Dobbin, R., 2004. Sex differences in voice onset time:  
587 A developmental study of phonetic context effects in british english. *J. of ASA*  
588 116 (2), 1179–1183.
- 589 Witten, I., Bell, T., 1991. The zero-frequency problem: Estimating the probabilities  
590 of novel events in adaptive text compression. *IEEE Trans. on IT* 37 (4), 1085–  
591 1094.
- 592 Xiao, J., Flandrin, P., 2007. Multitaper time-frequency reassignment for nonsta-  
593 tionary spectrum estimation and chirp enhancement. *IEEE Trans. on SP* 55 (6),  
594 2851–2860.