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

State-of-the-art automatic speech recognition (ASR) systems typically use a sliding window with a length of about 30 ms and a shift of about 10 ms to extract features such as Mel Frequency Cepstral Coefficients (MFCCs) from the acoustic waveform 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).

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

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

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

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

Acoustic information relevant to the identification of plosive sounds has been studied in the literature (O'Brien, 1993; Whiteside et al., 2004; McCrea and Morris, 2005; Jiang et al., 2006). Plosive consonants are produced by first forming a complete closure in the vocal tract via a constriction at the place of articulation, during which there is either silence or a low-frequency hum (called voicebar / prevoicing). The vocal tract is then opened, suddenly releasing the pressure built up behind the constriction. This opening of the vocal tract's airway is manifested acoustically by a transient and/or a short-duration noise burst. The duration of the interval between the release of the plosive and the beginning of voicing in the vowel is called the voice onset time or VOT. During this interval there is silence and/or noise caused by the release and/or aspiration noise. The VOT is one of the many acoustic cues for distinguishing plosives. The acoustic cues relevant to the articulation of a plosive can be related to manner (plosive, nasal, ...), place (bilabial, alveolar, velar, ...) and voicing (voiced, voiceless). A comprehensive discussion of these cues can be found in chapter 5 of Borden and Harris (1984) and we limit ourselves to an enumeration here. The *manner cues* for plosives include the presence of the silent region in the stop gap (obstruction phase), the rapid formant transitions and particularly a low locus frequency for the first formant F1, sudden energy change, release burst and aspiration. The *place cues* for plosives include the burst centre frequency (i.e. the main spectral peak of the turbulence occurring at the release), the locus frequency for the second and third formant transitions and the VOT. The *voicing cues* for plosives include the VOT, the presence of aspiration, the presence of an audible F1 transition, the intensity of the burst and the duration of the preceding vowel.

In this paper, we describe a VOT estimation algorithm using a high resolution signal analysis method which will better preserve timing information than MFCCs can. The next section is devoted to this signal representation, the reassigned time-frequency representation (RTFR). This representation allows well-separated impulses, cosines and chirps to be precisely located in time and in frequency. Because speech can to some extent be seen as a sum of such signals, we advocate the use of this representation for our current task. In section 3, the VOT characteristics are highlighted. A VOT estimation algorithm starts with indentifying segments of speech that potentially contain a plosive sound. We therefore describe our plosive data sets in section 4 and move on to section 5 where the actual algorithm that computes the VOT feature from the RTFR is described. Although the VOT has already been studied extensively, there are not many algorithms described to *automatically* extract this feature. Related work can be found in Lefebvre and Zwierzynski (1990); Ramesh and Niyogi (1998); Niyogi and Ramesh (1998); Sonmez et al. (2000); Kazemzadeh et al. (2006). However, to our knowledge this is the first time that the RTFR has been used to reliably extract the VOT feature. The performance of our algorithm is evaluated in section 6.1, while section 6.3 illustrates the modelling complexity as well as the usefulness of our automatic VOT extraction algorithm for phonetic studies by measuring some statistics of the VOT feature on the TIMIT database. Finally, in section 6.4 a rescoring approach shows a modest improvement in speech recognition accuracy using VOT. Conclusions can be found in section 7.

2 Spectral reassignment

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

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

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

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

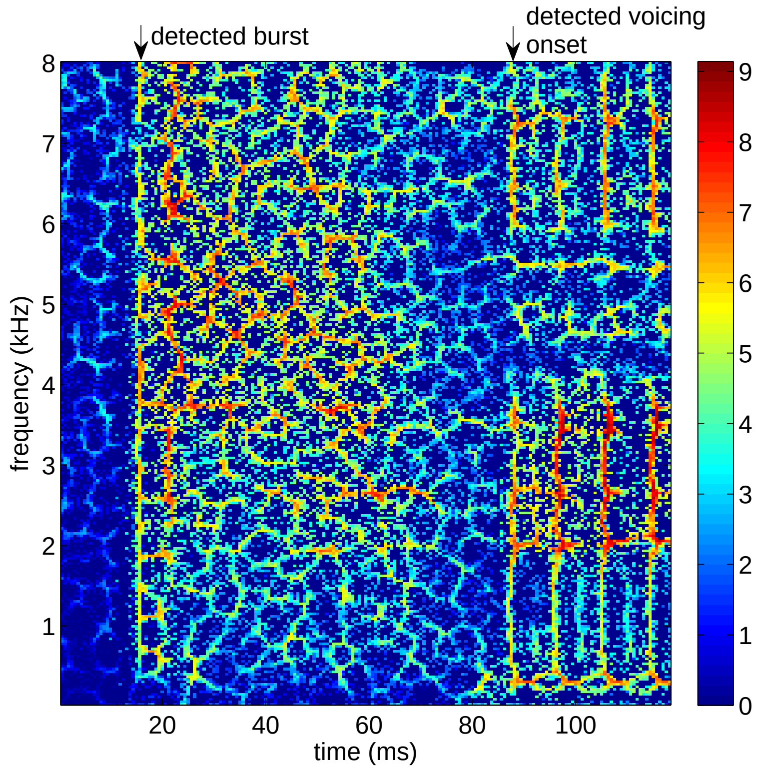

Fig. 1. *Reassigned time-frequency representation of a /t/ segment followed by /ih/. Colors encode the logarithm of the energy.*

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

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

3 Properties of the Voice Onset Time

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

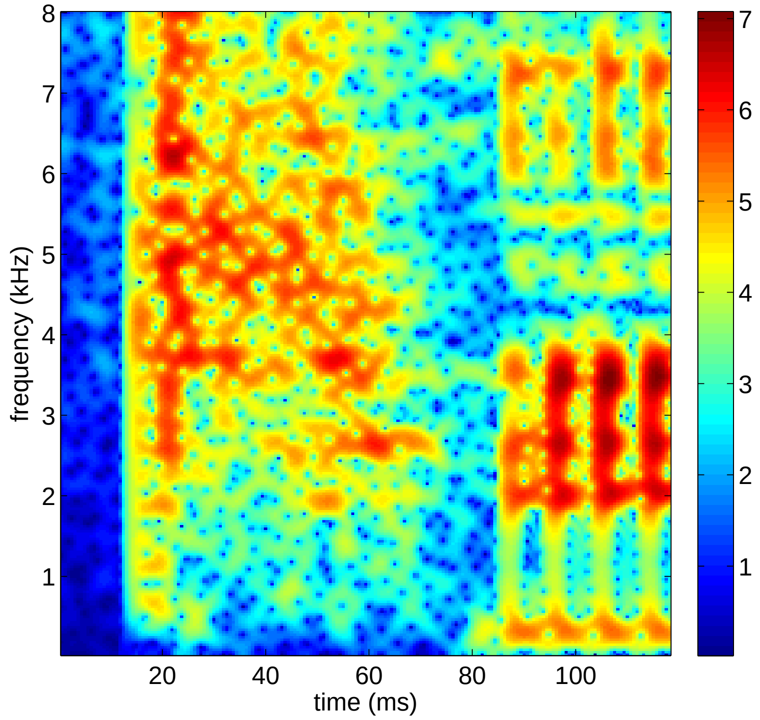

Fig. 2. *STFT representation of the /t/ segment from figure 1. Color encode the logarithm of the energy.*

$$\begin{array}{lll} \mathrm{VOT}(/b\,d\,g/) & < \mathrm{VOT}(/p\,t\,k/) & \\ \mathrm{VOT}(/b/) & < \mathrm{VOT}(/d/) & < \mathrm{VOT}(/g/) \\ \mathrm{VOT}(/p/) & < \mathrm{VOT}(/t/) & < \mathrm{VOT}(/k/) \end{array}$$

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

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

4 Data sets

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

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

4.1 The ”forced” data set

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

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

4.2 The ”free” data set

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

4.3 The ”manual” data set

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

burst or the start of voicing were removed. There is no constraint on the left phonetic context. Table 1 shows the exact number of examples thus retained in the ”manual” data set.

4.4 The ”test” set

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

5 The VOT estimation algorithm

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

The procedure has different possible outcomes. First, the plosive detection may fail by generating a false alarm or by missing a plosive. This leads to unrecoverable errors in the estimated VOT. Second, the generated segment boundaries may deviate too much from the real start or ending such that a different acoustic event is identified as the burst or voicing onset. For instance, if the proposed segment start is too early, an event belonging to the previous phone may be identified as the burst onset. If the proposed segment start is too late, the burst onset may not be detected. In the latter case, the missed event will be related to the segment boundaries proposed by the detector (see below), but given the erroneous segment boundary, the VOT error will be important. Third, either burst or voicing may not be revealed by their acoustic measure, in which case fallback estimates of their time of occurrence are derived from the segment boundaries proposed by the plosive detector. In this case, the VOT errors critically depend on the quality of the generated segment boundaries. Fourth, the segment may be correctly identified as a plosive with successful timings of the burst and voicing onset, leading to small VOT estimation error related to the time-frequency representation.

5.1 Detection of plosive segments

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

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

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

5.2 Burst onset detection

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

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

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

5.3 Start of periodicity

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

To obtain a robust estimate of the start of voicing, only the frequency range 0-4 kHz is retained. At a sampling frequency of 16 kHz as used in this work, this comes down to keeping only the lower half of the RTFR. Then, a short term autocorrelation is computed by multiplying every RTFR frame (for every 0.625 ms frame advance) with a weighted version of the frames at lags 1 to 40 and summing these values over the lag index and over the retained frequency bins. The weighting function (figure 3) is given by the difference of two decaying exponential functions and has a large value in the adult pitch period range of 5 to 20 frames, corresponding to a pitch period between 3.1 ms and 12.5 ms or a pitch frequency of 320 Hz down to 80 Hz. An asymmetric weighting function is chosen because we want to extract 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).

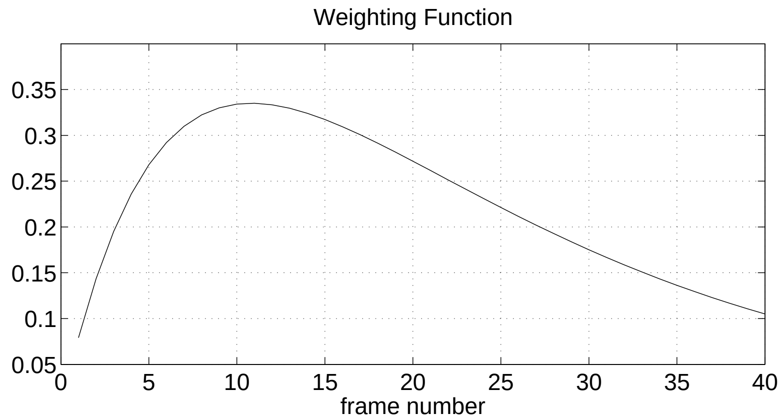

Fig. 3. *Weighting function of the periodicity detector.*

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

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

If the algorithm cannot detect voicing within the extended segment, the end of the unextended segment is marked as the start of voicing, i.e. we fall back to the HMM’s decision of the start of the next phone. This is a reasonable choice for English, where VOTs are mostly positive, but for other languages, voicing may already start in the closure interval. On the ”manual” data set, we measure a square root of the mean square error of 12.2 ms if voicing was detected, while it increases to 17.8 ms if voicing could not be detected within the extended segment. Not surprisingly, the HMM does a better job at detecting the start of the next vowel than it does at detecting the burst.

5.4 Discussion

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

6 Experiments

6.1 Algorithm performance for phonetic studies

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

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

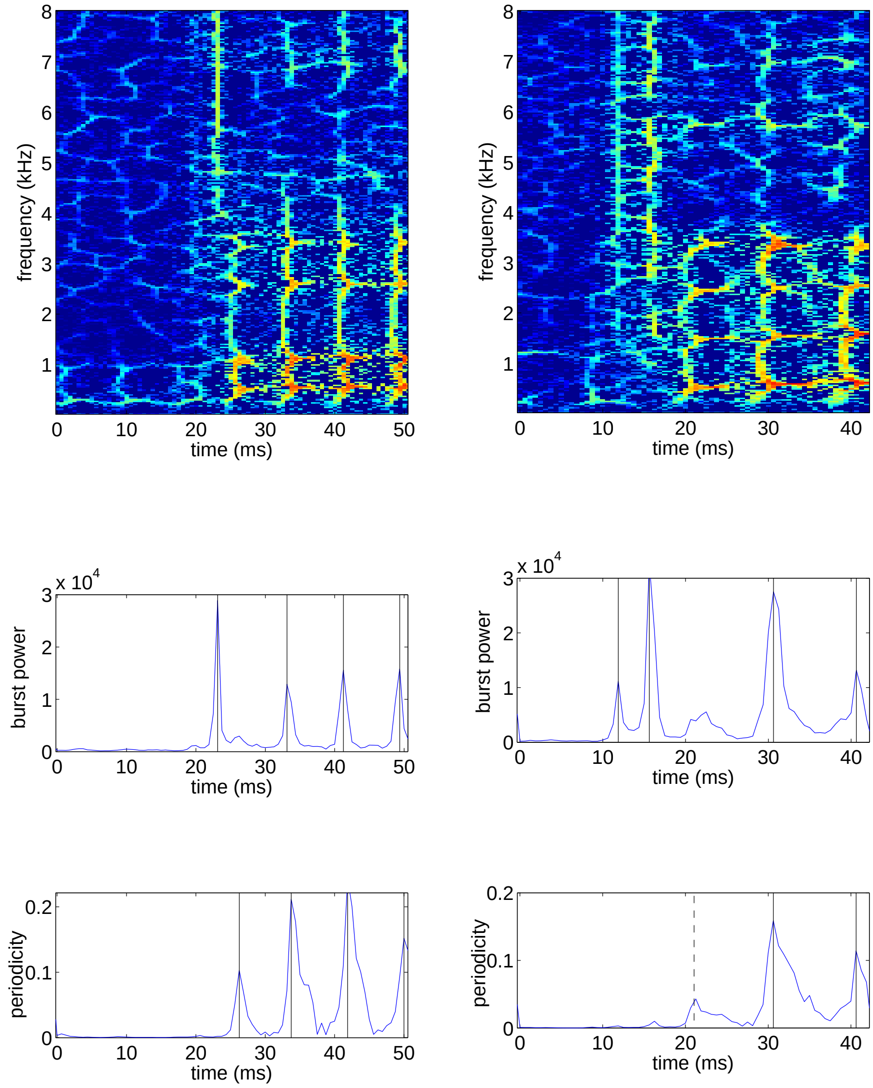

Fig. 4. *Left: illustration of the peak picking on a /b/ segment with a right context /aa/ (from "flat bottom") taken from the "free" data set. From top to bottom: RTFR, burst detection and periodicity detection. The peaks that satisfy the selection criteria are marked with vertical solid lines. Right: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

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

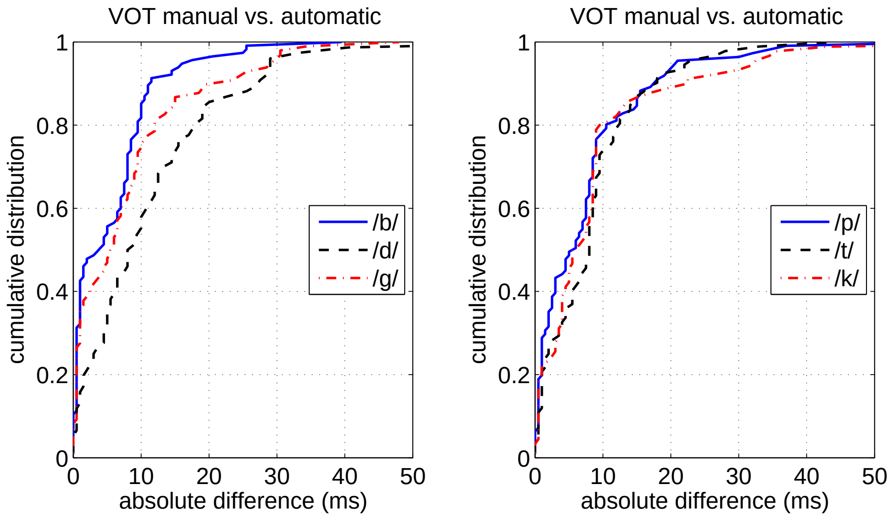

Fig. 5. *Absolute difference between the manually and the automatically extracted voice onset time.*

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) all segments				VOT (ms) 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

6.2 Algorithm performance for automatic speech recognition

While the above accuracy analysis is relevant for e.g. phonetic studies, where segment boundaries can be generated based on a manually produced phonetic transcription, its validity can be questioned in a fully automatic setting, where the goal of VOT estimation could be to improve speech recognition accuracy on plosives. Therefore, in the second study, the absolute difference between manual and automatic estimates is analysed on the "free" data set. However, an automatic phone 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

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

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

6.3 Estimated VOTs

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

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

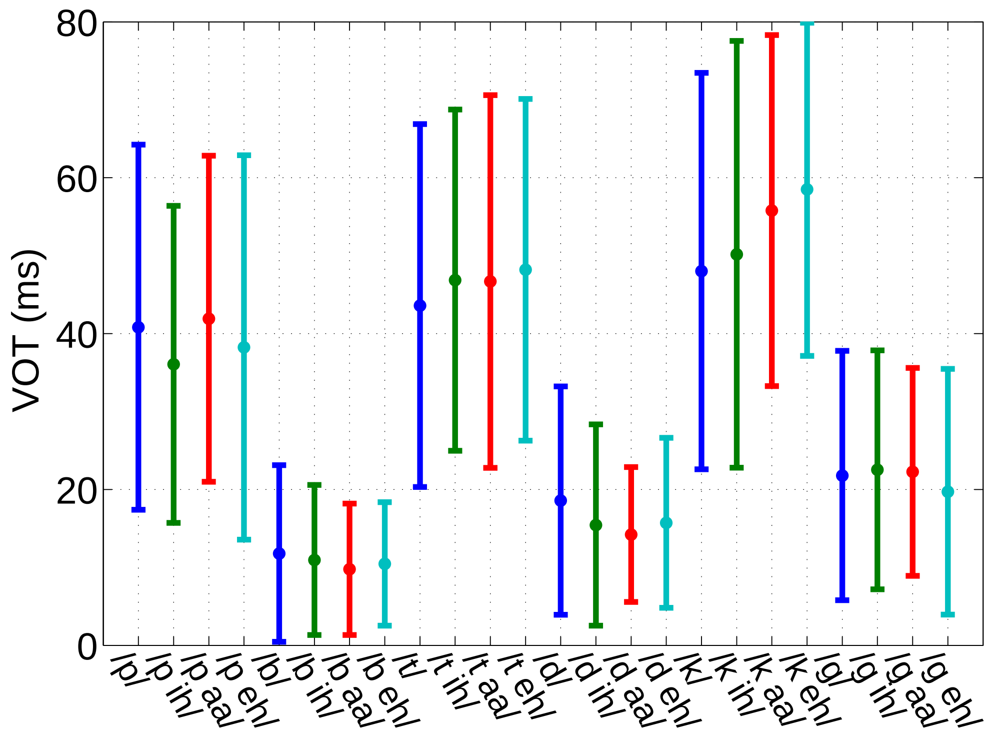

Fig. 6. *Mean VOT for plosives /p b t d k g/ by context (context independent, right context /ih/, /aa/, /eh/). The left context is always unconstrained. Error bars indicate +/- one standard deviation. Measured on the ”forced” data set.*

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

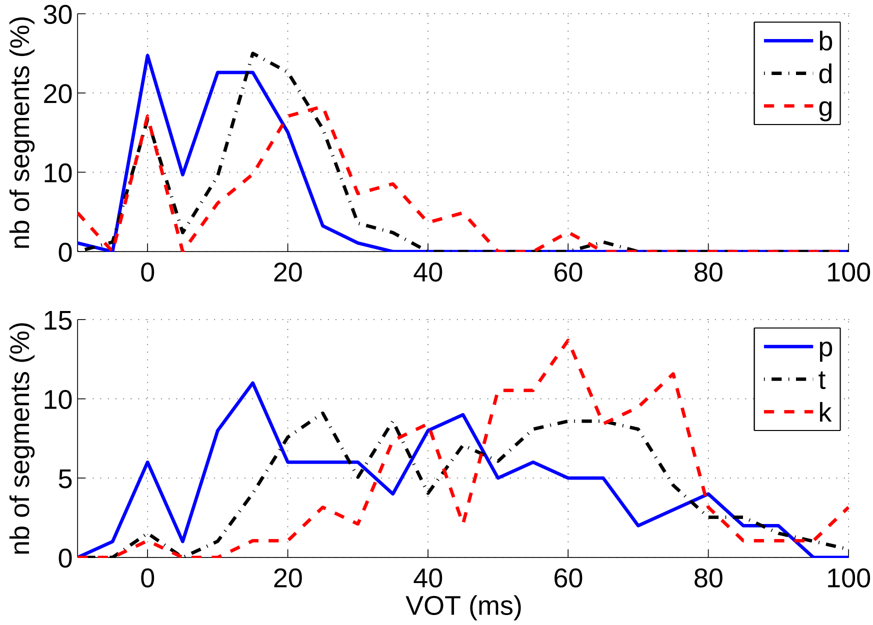

Fig. 7. *Normalised histogram of VOT estimates on the "forced" data set for plosives /b d g/ and /p t k/ followed by vowel /eh/, without constraint on the left context.*

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

6.4 VOT as a feature for automatic speech recognition

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

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

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

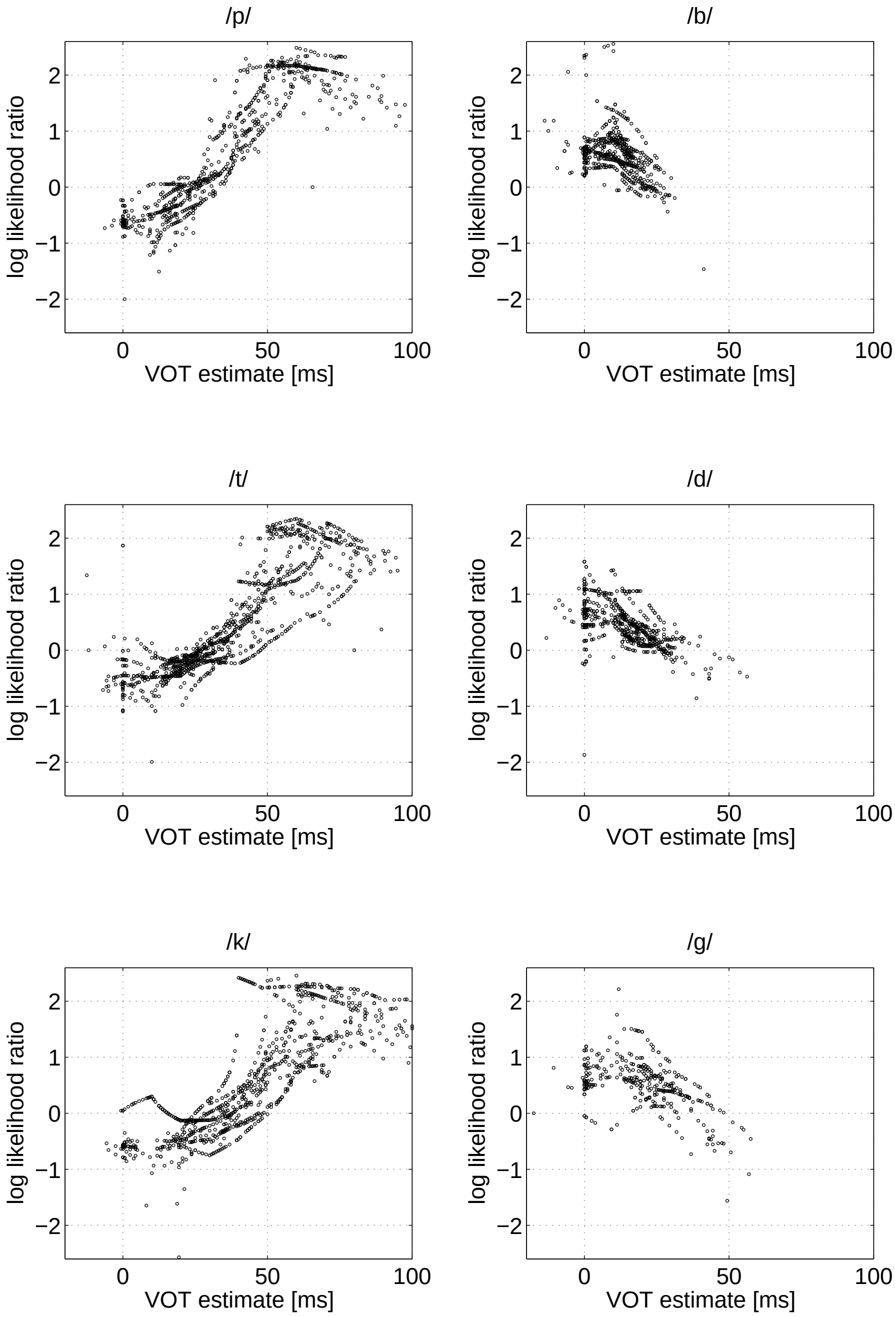

Fig. 8. *Logarithm of the likelihood ratio versus the automatically calculated VOT value, measured on the "test" data set.*

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

Figure 8 shows the logarithm (to base 10) of the likelihood ratio versus the estimated VOT value for the "test" data set. This set contains data that was not used during the construction of the histograms, while the ground truth about plosive identity and its context is known from the manual labeling provided in the TIMIT database. So let $P(V|l, p, r)$ be the probability that the estimated VOT falls in bin V for plosive p as measured on its histogram, and let $P(V|l, \overline{p}, r)$ be the probability read from the histogram for the plosive with opposite voicing. The log-likelihood ratio is then

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

where

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

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

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). These are the same sentences as used in the ”test” data set, but now the lattice will include more plosive candidates. The best path through the lattice will generate the phone segmentation of the ”test” data set. In formula 1, the likelihood $L(A)$ of each plosive arc A in the lattice is then linearly combined with the log-likelihood ratio of it being correct versus its variant with opposite voicing being correct. There is, however, a difference with the above. When dealing with the ”test” data set, the left and right phonetic contexts are unique. In a lattice, multiple arcs may arrive in the starting node of A and multiple arcs may leave from its ending node, so the left and right phonetic context are not unique. We denote the set of phone labels of arcs ending (or starting) in the starting (or ending) node of arc A with $\mathcal{L}$ (or $\mathcal{R}$) and sum the statistics over all contexts of A allowed by the lattice:

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

The corrected acoustic likelihood of a lattice arc A becomes:

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

Linear combination of log-likelihoods of different information sources was examined in Beyerlein (1998). The single free parameter α we introduced was tuned on the ”forced” data set, which is independent of the ”test” data set. This procedure reduced the phone error rate from 26.70% to 26.53% on the TIMIT test set. Hence, we observe that the VOT feature has contributed only very little to error rate improvement. This is not surprising, since we observe in figure 8 that the log-likelihood ratio can become negative for valid utterances of the plosive. On the other hand we have to realize that we attempt to correct only the plosive hypotheses generated by the HMM system, and this only along the voicing dimension. We can find 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. This yields an error rate floor of 25.85%. Hence, we have obtained $(26.7 - 26.53)/(26.7 - 25.85) = 20\%$ of the performance gain that would be achievable using an ideal voicing detector. In absolute numbers, the VOT-based likelihood ratio test corrected 80 out of 1853 plosive errors and hence the improvement is statistically significant. The gain shows that the VOT estimate does contain information that the HMM is not able to exploit. Apart from the overlap in the distributions of the VOT, the performance in this particular implementation is also limited by the pruning in the phone lattice. Each plosive hypothesis (arc) is rescored, but this can only lead to a change in decision if the hypothesis with opposite voicing is also in the lattice (and receives a better combined score). Hence, if the alternate, correct hypothesis was not included in the lattice because of pruning, it cannot be recovered, even with an ideal voicing detector. Further performance improvements might also be obtained by combining the HMM and VOT likelihoods in a non linear way.

7 Conclusions

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

8 Acknowledgement

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

References

Auger, F., Flandrin, P., 1995. Improving the readability of time-frequency and time-scale representations by the reassignment method. IEEE Trans. on SP 43 (5), 1068–1089.

Beyerlein, P., May 1998. Discriminative model combination. In: Proc. ICASSP. Seattle, WA, U.S.A., pp. 481–484.

Bilmes, J., Bartels, C., 2005. Graphical model architectures for speech recognition. IEEE SP Mag. 22 (5), 89–100.

Borden, G. J., Harris, K. S., 1984. Speech Science Primer: Physiology, Acoustics, and Perception of Speech, 2nd Edition. Williams & Wilkins, Baltimore, U.S.A.

Demuynck, K., Feb. 2001. Extracting, modelling and combining information in speech recognition. Ph.D. thesis, K.U.Leuven, Belgium.

Demuynck, K., Van Compernolle, D., Van hamme, H., Sep. 2006. Robust phone lattice decoding. In: Proc. ICSLP. Pittsburgh, U.S.A., pp. 1622–1625.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V., Oct. 1990. The DARPA TIMIT acoustic-phonetic continuous speech corpus. In: Speech Disc 1-1.1, NTIS Order No. PB91-5050651996.

Hainsworth, S., Macleod, M., 2003. Time-frequency reassignment: a review and analysis. Tech. Rep. CUED/FINFENG/TR.459, Cambridge University Engineering Department.

Jiang, J., Chen, M., Alwan, A., 2006. On the perception of voicing in syllable-initial plosives in noise. J. of ASA 119 (2), 1092–1105.

Kazemzadeh, A., Tepperman, J., Silva, J., You, H., Lee, S., Alwan, A., Narayanan, S., Sep. 2006. Automatic detection of voice onset time contrasts for use in pronunciation assessment. In: Proc. ICSLP. Pittsburgh, PA, U.S.A.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Comp. Speech and Lang. (14), 333–353.

Lee, C.-H., Clements, M. A., Dusan, S., Fosler-Lussier, E., Johnson, K., Juang, B.-H., Rabiner, L., Sep. 2007. An overview on automatic speech attribute transcription (asat). In: Proc. INTERSPEECH. Antwerp, Belgium, pp. 1825–1829.

Lefebvre, C., Zwierzynski, D., Nov. 1990. The use of discriminant neural networks in the integration of acoustic cues for voicing into a continuous-word recognition system. In: Proc. ICSLP. Kobe, Japan, pp. 1073–1076.

McCrea, C., Morris, R., 2005. The effects of fundamental frequency level on voice onset time in normal adult male speakers. J. of Speech, Lang. and Hearing Res. 48, 1013–1024.

Niyogi, P., Ramesh, P., May 1998. Incorporating voice onset time to improve letter recognition accuracies. In: Proc. ICASSP. Seattle, WA, U.S.A., pp. 13–16.

O'Brien, S., 1993. Spectral features of plosives in connected-speech signals. Int. J. Man-Mach. Studies 38, 97–127.

Plante, F., Meyer, G., Ainsworth, W., 1998. Improvement of speech spectrogram accuracy by the method of reassignment. IEEE Trans. on SAP 6 (3), 282–286.

Ramesh, P., Niyogi, P., Nov. 1998. The voicing feature for stop consonants: Acoustic phonetic analyses and automatic speech recognition experiments. In: Proc. ICSLP. Sydney, Australia.

Seppi, D., Falavigna, D., Stemmer, G., R., G., Sep. 2007. Word duration modeling for word graph rescoring in lvcsr. In: Proc. INTERSPEECH. Antwerp, Belgium, pp. 1805–1808.

Sonmez, K., Plauche, M., Shriberg, E., Franco, H., Oct. 2000. Consonant discrimination in elicited and spontaneous speech: a case for signal-adaptive front ends in ASR. In: Proc. ICSLP. Beijing, China.

Stouten, F., Martens, J.-P., Sep. 2006. Speech recognition with phonological features: Some issues to attend. In: Proc. ICSLP. Pittsburgh, PA, U.S.A., pp. 357–360.

Whiteside, S., Henry, L., Dobbin, R., 2004. Sex differences in voice onset time: A developmental study of phonetic context effects in british english. J. of ASA 116 (2), 1179–1183.

Witten, I., Bell, T., 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Trans. on IT 37 (4), 1085–1094.

Xiao, J., Flandrin, P., 2007. Multitaper time-frequency reassignment for nonstationary spectrum estimation and chirp enhancement. IEEE Trans. on SP 55 (6), 2851–2860.