DOI: 10.17469/O2104AISV000016

DUCCIO PICCARDI, FEDERICO BECATTINI

Voice Onset Time Enhanced User System (VOTEUS): a web graphic interface for the analysis of plosives' release phases

The paper proposes an up-to-date literature review of the works using AutoVOT, a discriminative large-margin learning algorithm developed for the semi-automatic measurement of voice onset times. In order to expand the accessibility of the tool in linguistic research, we present VOTEUS, a user-friendly graphic interface written in Python. The interface is conceived to assist the researcher throughout the whole process of annotation, from the forced alignment of the corpora to the refinement of the AutoVOT tier and the extraction of the durations. The general aim is to speed up this phase of data analysis, providing a significant improvement on prevalent practice to date.

Keywords: AutoVOT, Graphic User Interface, VOTEUS, annotation, forced alignment.

1. *Voice Onset Time: Tools for a middle-aged feature*

It has been roughly fifty years since the first description of the Voice Onset Time (VOT) as «the interval between the release of the stop and the onset of glottal vibration, that is, voicing» was proposed in Lisker, Abramson (1964: 389). Celebrating this anniversary, Abramson and Whalen (2017) wrote a retrospective essay[1] discussing the evolution of its denotation and some critical points, not without proposing recommendations on Praat (Boersma, 2001) tiers labeling in VOT research. In the last paragraph, the authors give some space to a brief recollection of tools developed for the automatic measurement of VOT, hoping that «these systems will continue to improve in the coming years» (Abramson, Whalen, 2017: 84). Particular attention is dedicated to AutoVOT, described as «the most widely used system» (ibid.: 83). In the following sections, we will describe AutoVOT and provide an up-to-date account of its applications, highlighting the necessity to broaden its audience. We will then introduce Voice Onset Time Enhanced User System (VOTEUS), a web-based interface currently in development that will facilitate the usage of AutoVOT, also integrating other functionalities for VOT annotation.

[1] Abramson, Whalen (2017) leads the way to a special VOT issue of the 2018 *Journal of Phonetics*, dedicated to theoretical and experimental aspects of voicing contrasts (Cho, Docherty & Whalen, 2018).

1.1 AutoVOT: Procedures and performances

AutoVOT is a discriminative large-margin learning algorithm for VOT semi-automatic measurement originally developed by Morgan Sonderegger and Joseph Keshet (2012)[2], and subsequently integrated in a software written in Python and based on declarative programming (Keshet, Sonderegger & Knowles, 2014). While first thought for the annotation of voiceless plosives, AutoVOT expanded its agenda to work with prevoiced plosives (i.e. negative VOT; Henry, Sonderegger & Keshet, 2012) and preaspirated plosives (Sheena, Hejná, Adi & Keshet, 2017). AutoVOT is compatible with *.wav files (16 kHz mono) and Praat textgrids (*.TextGrid). The algorithm can be used to train models providing *.wav files with hand-measured textgrids containing a common label (e.g. *vot*) as input. The trained model can later be applied to new *.wav files matched with textgrids structured with a tier with aligned intervals in order to segment the contained VOTs. The recommended, optimal tiers should not include more than one stop consonant and should begin 50 ms before the stop burst or 30 ms before the entire segment. Eventually, AutoVOT predictions should be checked and adjusted by a human annotator. Among the studies that made use of AutoVOT (see below), few actually reported precise information on its performance. In this aspect, Stuart-Smith, Sonderegger, Rathcke & Macdonald (2015) provides the most comprehensive picture and will be here summarized as an example of the algorithm's potentials and problematics.

1) The aligned tiers were automatically generated, and not subsequently modified;
2) the authors provided two different small training sets (100 VOTs from each of five analyzed speakers per set) to generate a model for voiceless plosives and one for voiced plosives;
3) after the application of the models to the inquired corpus, the phase of evaluation and correction of AutoVOT predictions had an astounding 1:1 ratio between the actual duration of the annotated file and the time of human adjustment;
4) the variable ANNOTATOR in the statistical analysis did not hold significance, hinting to the good quality of the semi-automatic measurement;
5) the miscellaneous quality of the recordings contained in the inquired corpus did not alter the effectiveness of the algorithm;
6) a total 2564 predictions were labeled as "not usable" (21,6%; 1736 voiced stops, i.e. 29,8% and 828 voiceless ones, i.e. 7,9%), 5860 as "correct" (62,6%; 3171 voiced stops, i.e. 54,4% and 2689 voiceless ones, i.e. 76,2%) and 1474 as "corrected" after the phase of human adjustment (15,8%; 916 voiced stops, i.e. 15,7% and 558 voiceless ones, i.e. 15,8%)[3].

[2] A first attempt by the two authors to tackle the issue of automatic VOT measurement can be read in Sonderegger, Keshet (2010).

[3] The reported percentages refer to parts of the total number of analyzed tokens (9898; 5823 voiced and 4075 voiceless stops; see below). A prediction was coded as "not usable" in the case of alignment or transcription errors, sounds overlapping to the token production or variation phenomena hindering

1.2 Literature review

In this section we will provide a review of all the linguistic research and ongoing projects[4] reporting the use of AutoVOT and indexed as such in Google Scholar[5]. Starting with studies on corpora of read speech, Chodroff, Godfrey, Khudanpur & Wilson (2015) applied AutoVOT on a total of 68000 tokens produced by 129 American speakers. The authors searched for /b d g p t k/ VOT variability in a corpus thought to be quantitatively appropriate for observations at both the talker-specific and the population level. Results showed indeed significant differences in individual productions, such as the entities of the effect of stop category or speech rate on VOT lengths. The authors also found that the individual, within-category durational means and standard deviations were consistently connected, and that VOT lengths were strongly correlated across the stop categories elicited in individual productions, pointing to structured variability of VOT patterns[6].

Bang, Sonderegger, Kang, Clayards & Yoon (2018) explored the topic of a sound change regarding Seoul Korean aspirated plosives through the analysis of 6849 intonational phrase-initial stops elicited by 118 speakers and contained in an apparent-time corpus[7]. The study confirms the previously retrieved distribution showing that the female speakers are leading the substitution of VOT length with f0 patterns as primary phonological cue of the aspirated series; moreover, the change has slowed down in recent years, hinting to its near completion. The frequency of a word is positively correlated with both the degree of VOT reduction for aspirated plosives and f0 contrast enhancement; since this last result is contra-

---

the realization of the token as a plosive.

[4] From the moment that the main goal of this section is to create a broader understanding of the potentialities of the tool in actual linguistic inquiries, we will exclude from the review the project papers stating the intent to integrate AutoVOT in other tools for linguistic analysis, such as McAuliffe, Stengel-Eskin, Socolof & Sonderegger (2017). In the review, we will focus our attention on the results of the experiments concerning VOT measures, with the *caveat* that VOT is not always the only, nor the main feature analyzed in the reported research.

[5] We are well aware that the transparency practices concerning the use of software in linguistic research are not homogeneous among the different subfields of the discipline. However, phonetics reportedly values the quotation of the equipment (Berez-Kroeker, Gawne, Kung, Kelly, Heston, Holton, Pulsifer, Beaver, Chelliah, Dubinsky, Meier, Thieberger & Woodbury, 2018: 9-10) so that we hope that our search will result exhaustive.

[6] Chodroff, Wilson (2017) confirmed these results comparing the data with hand-labelled productions in a laboratory setting, while increasing the number of tokens annotated with AutoVOT (88725 tokens, 180 speakers). In this study, the authors present an extensive discussion on the implications of these outcomes for structural constraints on phonetic systems and perceptual adaptation. Finally, Chodroff, Wilson (2018) replicated the findings through the semiautomatic annotation of 96357 VOTs from the same corpus, also describing a similar structured variability for plosives' center of gravity and onset f0 in the following vowel.

[7] Partial results from this study (5888 tokens) can be found in Bang, Sonderegger, Kang, Clayards & Yoon (2015). In Bang (2017) the data is further compared to corpus-retrieved American English (126 speakers, 4208 tokens) and German (118 speakers, 2660 tokens) read speech.

ry to typological expectations, the authors suggest that the f0 enhancement is an adaptive change to the VOT reduction. On the other hand, the presence of a subsequent high vowel inhibits these processes, suggesting that general coarticulatory lengthening mechanisms could have conditioned the modalities of the change. Finally, intrinsic f0 vowel differences after voiceless plosives are dampened as the phonological f0 distinctions arise over time. These results are discussed in relation to a potential Seoul Korean tonogenesis.

The same research topic is investigated in Cheng (2017) from the point of view of South Californian Heritage Korean. 32 speakers were recruited to read a set of 35 words presenting fortis/lenis/aspirated stops and affricates minimal pairs. Participants were classified in three generational levels, corresponding to different times and modalities of exposure to Korean and American English. The set was read in a fixed carrier phrase and in more naturalistic sentences, resulting in a total of 2240 tokens. The tonogenetic shift was well represented by the first-generation speakers, while the second-generation ones seemed to use primarily VOT lengths to express phonological distinctions. Still, a perceptual counterpart focused on language attitudes showed that this difference alone cannot be considered a marker of linguistic proficiency for South Californian Koreans. The results are compared to similar tendencies described in other studies and interpreted in the light of potential attrition with American English.

Together with this last study, Schertz, Kang, & Han (2017) is of particular interest to the aims of this section for successfully applying AutoVOT to different consonantal typologies. The authors gathered 11121 productions of Korean and Mandarin sibilants and affricates from an isolated-words reading task proposed to 107 bilingual speakers from the two Chinese prefecture-cities of Hunchun and Dandong, located at the border with North Korea. After analyzing the VOTs of all the tokens for the phonetic description of the phonological categories of the two coexisting systems, a subset of corresponding sounds is further compared to observe the intertwined participation of the two languages to potential sound changes. In regard of the corresponding affricates, results show that older interviewees equate the Korean and Mandarin VOT values, while peculiar trends can be observed in their younger counterparts. In Dandong, young speakers present a Seoul-like tonogenetic tendency in their Korean production, leaving the Mandarin tokens unaffected. In Hunchun, this demographic group has shorter VOTs in both Korean and Mandarin productions; however, no VOT merger is observable in the Korean phonological categories, being the change probably Mandarin-driven.

Singh, Keshet, Gencaga & Raj (2016) tackled the debated issue of VOT-physical age patterns, grounding their results on unprecedented quantities of tokens[8] and observed speakers (630, American English). The authors make a successful use of AutoVOT in also predicting the Voice Offset Times contained in the corpus, i.e.

[8] It should be noted that the exact number of VOTs is not stated in the paper. The authors report that all the stop plosives (/b d g p t k/) were represented at least once for each speaker in the corpus. We can infer that the study applies AutoVOT *at least* to 7560 tokens.

«the duration between the cessation of voicing in a voiced phoneme, and the onset of the burst of the subsequent plosive sound» (ibid.: 2). Contrarily to previous results, both these features do not show significant correlations with speakers' age. The accuracy of the annotation method is believed to be an important factor in determining the outcome of the research.

Chen, Xiong & Hu (2018) preferred a real-time approach to this same research topic. The authors extracted 1001 voiced and 2297 voiceless plosives from 40 recordings of the Christmas speeches by Queen Elizabeth II ranging from 1953 to 2016. While controlling for potentially conditioning linguistic factors, the researchers observed a declining trend in the amplitude of annual fluctuations in VOT productions, parallel to a similar tendency in VOT mean values. These findings are tentatively interpreted as dependent from physiological factors of vocal aging.

Goldrick, Keshet, Gustafson, Heller & Needle (2016) studied VOT durations of the slips of the tongue occurring during tongue twisters. 34 American English speakers were invited to read the materials in time to a metronome. The twisters were composed by monosyllabic stimuli selected to differ just by the sonority of the first plosive, with four different typologies (ABBA, BAAB, ABAB, BABA). 68000 tokens were segmented with AutoVOT. The slip of the tongue was defined as a deviation from a normal VOT duration in the direction of the other member of the twister (e.g., /b/ with long VOTs and /p/ with short releases). The switching patterns (ABBA, BAAB) showed errors with smaller degree of VOT deviations than those from the alternating ones. Moreover, erroneous productions had a higher degree of variation than the correct ones. The authors discuss their acquisitions in the light of the two proposed explanatory factors for this kind of speech errors, i.e. planning and articulatory processes, finally suggesting an integrated account.

Coming to the analyses of large corpora of spontaneous speech, Stuart-Smith, Sonderegger, Ratchke & Macdonald (2015)[9] studied 9898 voiceless and voiced[10] plosives uttered in Glaswegian vernacular by 23 working-class women. Two clusters of recordings were taken into account, one from the 1970s and the other from the 2000s; three age groups were established per cluster, searching for proofs of a historical process of VOT lengthening in the inquired variety. Elderly speakers from the 1970s had significantly shorter VOTs than their younger counterparts; in the 2000s, the situation is reversed, with the least pronounced aspirations uttered by the youngest speakers. These two different directions underline a sociophonetic potential of VOT related to speakers' age in Glaswegian. Moreover, the fact that middle-aged and old speakers from the 2000s showed longer VOTs than their respective age groups from the 1970s seems to suggest a real-time lengthening. The aberrant results from the youngest group from the 2000s is tentatively explained in

[9] Stuart-Smith, Ratchke, Sonderegger & Macdonald (2015) reported preliminary results from 12 speakers and 6125 tokens, reduced to 3012 reliable measures.

[10] As of 2015, the algorithm for negative VOTs (Henry et al., 2012) did not prove to be reliable; as consequence, all the voiced plosives observed in this study had positive VOTs. This problem seems to be somehow resolved at the time of Solanki (2017) (see below).

reference to a reported tendency of this demographic cluster to follow vernacular patterns of speech: in this case, Scots is known for its short VOTs, whereas Scottish Standard English has longer values, more similar to Anglo-English.

Sonderegger, Bane & Graff (2017)[11] took 25584 VOT durations from the speech of twenty participants to a British reality television show produced over more than fifty consecutive days. The research goal was to describe individual speech dynamics in medium term. Time dependence was pervasive in the productions of all the participants; in particular, by-day variability was the norm, while time trends interested around half of the observations. This result somehow conciliates the concepts of individual dynamicity in short-time and individual stability in long-time, from the moment that not all the daily fluctuations have the potential to become consistent change. Moreover, the study bore little evidence of overall convergence over time between the productions of the participants, challenging the assumptions of change by accommodation theories. However, a consistent convergence between the values of two participants after their romantic engagement was observed, hinting to the fact that strong social bonds represent a determinant factor in such dynamics. Finally, the participants showed different estimates of phonetic plasticity: the authors suggest that this parameter is central in determining the role of a speaker as innovator or early adopter in language change.

Two Ph.D. dissertations cited AutoVOT for the segmentation of semi-spontaneous productions in laboratory tasks. Turnbull (2015) used the algorithm to segment 1748 VOT tokens derived from an experiment with 19 participants, in the framework of listener-oriented accounts of predictability-based phonetic reduction. The participants sat in front of a screen, that showed highlighted words beginning with a plosive. The task was to instruct a confederate to click on the same word on his other screen, without the possibility of directly viewing it but being instructed that the two supports were showing the same elements. The researcher tested for the effect on VOT length of stop place, context (the fact that the word had no minimal pairs, or had minimal pairs without competitors on screen, or had minimal pair with competitors on screen), phonological neighborhood density, log frequency and individual scores to assess the extent of the Theory of Mind of the participants. Surprisingly enough, only the place of articulation had a significant effect, probably due to the choice of the experimental materials.

Solanki (2017) studied speech accommodation in live conversation in a laboratory setting. 12 female participants from Glasgow were recruited and paired to verbally interact in front of two separated screens with the aim of finding a number of small graphic differences between elements placed in three different scenarios. A total of 14494 (negative and positive) VOTs was retrieved during these sessions. Results showed that neither the previous production of the interlocutor, nor the position of the interaction in the course of the experiment had a significant effect

[11] Preliminary VOT results can be found in Bane, Graff & Sonderegger (2010) (circa 800 manually segmented VOTs), Sonderegger (2012) (6494 tokens, from manual annotations and automatic measurements) and Sonderegger (2015) (20822 tokens analyzed with AutoVOT).

causing convergence. However, interaction length proved to be a significant factor in VOT accommodation. The author infers that, while the time of communicative contact does not imply phonetic convergence *per se*, the content of the contact is crucial in assessing the will of cooperating. In this case, longer interactions were a signal of more difficult tasks, triggering convergent behaviors.

Finally, two research projects are planning to implement AutoVOT for the analysis of spontaneous productions. Chen, Kozbur & Yu (2015) transcribed fifteen years of oral arguments (1998-2013, 975 hours of recordings) that took place at the U.S. Supreme Court for the sake of analyzing speech accommodation phenomena. While preliminary results are available for vowel formants, the authors will also check for convergence in VOT values. Singh, Raj & Gencaga (2016) lists the voice onset time among those "stable" sub-phonemic features that could be of help in the field of forensic anthropometry from voice. The idea is to automatically extract VOT values from short audio segments involved in criminal activities, such as hoax calls (Singh, Keshet & Hovy, 2016), to infer physical features of the culprit facilitating the process of profiling.

1.3 Discussion

The proposed review highlights the versatility of the tool, that proved its usefulness in disparate research conditions (from real-time to apparent-time corpora, from lab speech to spontaneous and read speech), for very diverse corpus dimensions (from 1748 to 96357 VOTs) and topics of investigation (individual differences, sociophonetic values, interactional processes etc.). A concept recurring in the summarized studies is that automatic segmentation procedures are a key factor for exploring new nuances of the VOT feature, and grounding previous results on more adequate quantities of observations. However, in six years since Sonderegger, Keshet (2012), the algorithm was adopted in just 13 research projects, including Ph.D. theses and ongoing works. In addition to that, the direct involvement in 6 of these research of one of the original authors of AutoVOT definitely catches the eye[12]. Among the many factors that could explain these numbers, our take is that the level of accessibility of the technology at hand should not be taken lightly, especially in a field that dwells in deeply-rooted interdisciplinarity. Linguists have to master a wide variety of competences, both humanistic and scientific. The lack of expertise in one of its essential components results in being detrimental to the field itself[13], e.g. precluding the access to convenient tools. One possible solution resides in the development of user-friendly graphic interfaces apt to lighten the burden of specific tasks on research projects. The academic community is already working in this direction, with the planning of web-based interfaces such as DARLA (Reddy, Stanford, 2015) for semi-automated forced alignment and vowel extraction. In particular, the project

[12] In line with this fact, Solanki (2017) was written under the supervision of Stuart-Smith at the University of Glasgow.

[13] On this topic, see e.g. the informal sarcasm chosen by Foulkes (2015) to describe sociophonetics at the *ICPhS* dedicated session.

of Visible Vowels (Heeringa, Van de Velde, 2017), an internet tool for vowel plotting, normalization and analysis of dynamic features, puts great emphasis on its user-friendliness and accessibility (ibid.: 4034)[14]. It is from these premises that we present here Voice Onset Time Enhanced User System (VOTEUS), an open-access framework for semi-automatic VOT annotation of speech corpora.

2. *VOTEUS*

The framework Voice Onset Time Enhanced User System (VOTEUS) that we present in this paper is a tool intended to bridge the gap between the linguistic and computer science knowledge domains in the usage of AutoVOT. Our goal is to provide an intuitive interface to configure and run the algorithm on large speech corpora, without requiring any computer programming skills and therefore allowing everyone to exploit all AutoVOT's capabilities. More than that, we provide a set of functionalities that guide the user to easily generate fully annotated VOT datasets starting from raw speech recordings and their transcriptions. In particular we integrate a forced alignment routine to provide initial speech segments on which AutoVOT can be applied and we developed an intuitive interface within VOTEUS to manually refine the predicted VOT tiers, in order to produce high quality annotations. Furthermore, VOTEUS has been developed keeping in mind a modular software structure. This allows it to be extended and integrated with additional methods for detecting VOT and possibly compare them with AutoVOT. In the following we provide an overview of VOTEUS' architecture and organization and explain its main use cases, namely corpus inspection, semi-automatic annotation and training AutoVOT models. VOTEUS is currently under development for Linux and Windows operating systems and is going to be released under the MIT license, therefore allowing users to include and modify its source code within other projects. An alpha release of VOTEUS is scheduled to be released in early 2019. Code and installation guide will be available for download at the following link: https://github.com/fedebecat/VOTEUS.

2.1 Architecture

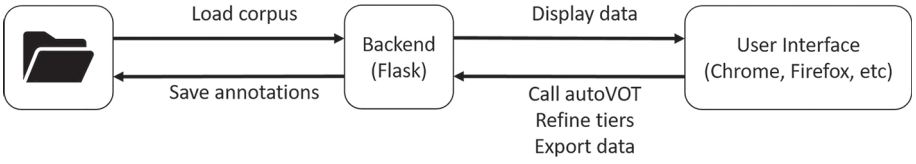
VOTEUS is organized into two software components: a backend that integrates and extends the functionalities of AutoVOT and the actual graphical user interface (GUI) for controlling and invoking these functionalities. This architectural choice keeps the control logic separate from the user interface, which reflects in a better code maintenability and implements the concept of separation of concerns by encapsulating different logical modules into separated software components[15]. We de-

[14] This point was also firmly made during a presentation of Visible Vowels held by Van de Velde at the Scuola Normale Superiore of Pisa (28/04/17).

[15] The concept of Separation of Concerns was initially introduced in Dijikstra (1982) and nowadays is at the basis of the most diffused architectural design patterns such as Model-View-Control (MVC).

veloped our backend framework in Python, by creating wrappers for AutoVOT and integrating them into a Flask[16] webserver which exposes the GUI in the form of an interactive web page, that runs in the browser. We adopted a web-based solution developed in Python more than integrating our system into existing tools or external engines such as Praat (Boersma, 2001) to focus on portability and diffusion among inexperienced users. Moreover, the advantages of this choice are twofold, in the one hand we provide a familiar environment to the user, minimizing the cognitive burden required to learn how to utilize the system, on the other we could exploit the vast resource of available libraries and toolkits for web development available online. The resulting user interface has been developed in JavaScript, largely exploiting jQuery[17] and the wavesurfer.js[18] and Google Chart Libraries[19]. All styling materials have been taken from the resources of materializecss[20]. Whereas our framework is developed as a web-based interface, we propose VOTEUS as a standalone application to be run on personal computers/workstations and not as a remote application accessible online, since users would need to upload and store large amounts of data for their corpora. At the same time it should be noted that remote access to interact with a VOTEUS instance could be easily enabled. To maintain compatibility with others systems we rely on the same data representation formats used by AutoVOT, in particular we store all audio annotations in textgrid files. VOTEUS is thought to handle different speech datasets that can be added to the framework simply including a folder to its search path. Also in this case we refer to AutoVOT specifications for input files (see above). In Figure 1 a schematic representation of the main modules and functionalities of our system is shown, depicting how the interface interacts with the data through the backend.

Figure 1 - *Schematic representation of VOTEUS' architecture. Data is stored on the disk and read by the backend. The interface allows the user to browse the data and call the functions exposed by the backend. The generated annotations are then saved back on the disk by the backend*



[16] Flask is a Python based micro-framework for developing web applications (http://flask.pocoo.org/).
[17] https://jquery.com/.
[18] We used the wavesurfer.js library (https://wavesurfer-js.org/) in combination with the spectrogram plugin (http://wavesurfer-js.org/example/spectrogram/).
[19] https://developers.google.com/chart/.
[20] http://materializecss.com/waves.html.

2.2 Corpus inspection

The simplest functionality offered by VOTEUS is to display large speech corpora in an aggregated fashion, in order to interactively inspect all the available annotations. Figure 2 shows how this is presented to the user through the interface. Once a corpus is loaded in the interface, the user can navigate through all the *.wav files and study their waveforms and spectrograms. A temporal representation of all the available tiers in the textgrid annotation file is shown and the user can highlight the correspondent interval in the audio representation (waveform and spectrogram) by simply clicking on the tier of interest. If multiple types or tiers are present in the textgrid, they are stacked inside the interface and color coded for simple inspection. Figure 3 depicts three different details of the interface, showing how the user can interact with the annotations by clicking on the tiers. The selected audio file can also be reproduced, both in its entirety or focusing on specific tier intervals. For long audio files, the waveform and spectrogram can be zoomed-in and out to examine the details of the recording at a fine-grained level.

Figure 2 - *Main Graphical User Interface of VOTEUS. When a corpus has been loaded, the user can browse its files and display the corresponding waveform and spectrogram. All available annotations are shown in the timelines below. The user can interact with the tiers to highlight or listen specific audio segments. The buttons in the lower part of the GUI can be used to call some of the functionalities of the backend.*

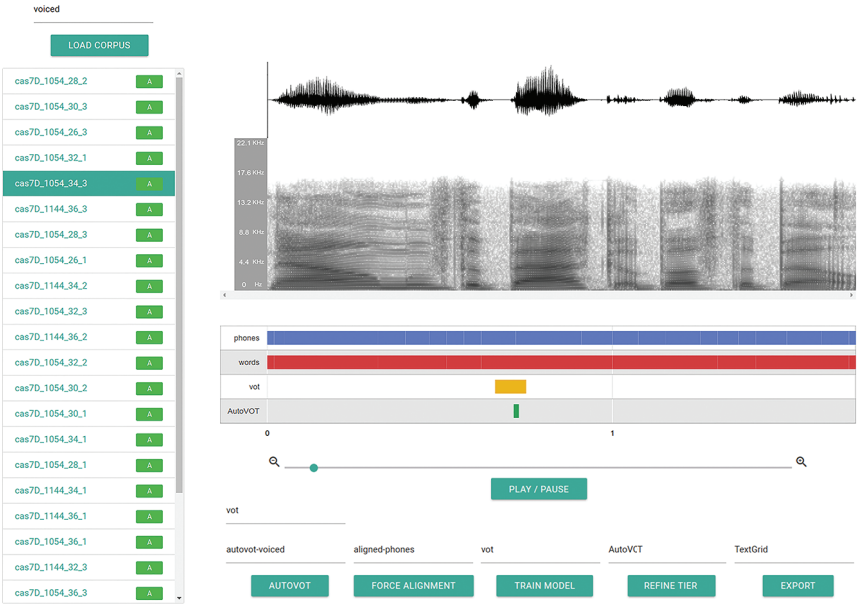
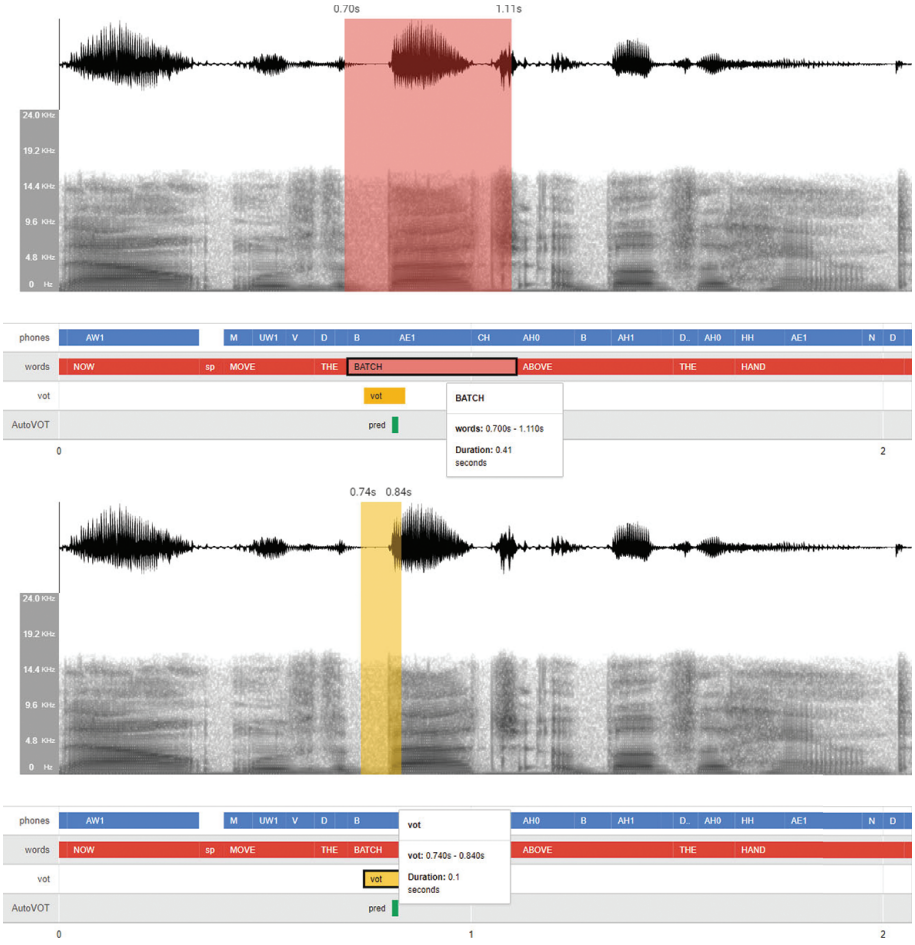


Figure 3 - *By interacting with the annotations, the user can isolate the interested portion of the audio file and reproduce it. Different tiers are highlighted with different colors*



2.3 Semi-automatic annotations

The most important feature provided by VOTEUS is the possibility of generating semi-automatic annotations of speech corpora for VOT intervals. This functionality is articulated into three distinct steps:

a) Automatic forced alignment
b) Fast refinement of textgrid tiers
c) Batch annotation with a pretrained model

To generate the annotations we rely on an AutoVOT model, which can be applied on text segments to generate VOT predictions. Whereas this process is fully automatic, it requires as input a collection of candidate speech intervals that should contain no more than one stop consonant and start 50 ms before the stop burst or 30 msec before the entire segment. To provide such segments we rely on SPPAS (Bigi,

2015; Bigi, Meunier, 2018), an additional tool that performs automatic forced alignment. The term forced alignment denotes a process for determining the time segment of a recording that contains a given portion of a transcription. SPPAS aims at automatizing this process to produce annotations with a granularity that ranges from utterance to phoneme. To this end, SPPAS performs three sub tasks divided into *tokenization*, *phonetization* and *time-alignment*. Tokenization (text-normalization) converts input text into a linguistic representation with standardized and ordinary words, phonetization applies a grapheme-to-phoneme translation and finally time-alignment deals with aligning the sequence of phonemes to the speech signal. SPPAS is provided with resources for multiple languages[21], but the authors state that most of the algorithms have been developed to be as much language-independent as possible and that adding a new language reduces to integrating a few resources such as lexicons and dictionaries. This aspect of SPPAS, in combination with the ready-to-use Python bindings for automatic phonetic segmentation, is what motivated our choice towards this tool. We wrapped SPPAS inside VOTEUS' backend and it can be easily invoked by the interface to obtain candidate intervals on which to apply AutoVOT. Since AutoVOT input requirements are quite strict, to provide better search intervals we implemented a fast refinement procedure for allowing the user to modify existing tiers[22] or adding new ones. By opening this view, VOTEUS shows in a rapid sequence all the annotations for the selected tier for each audio file in the corpus. The user can examine and click directly on the spectrogram to define the precise boundaries of the interval and move to the next annotated entry (Figure 4). If any modification is made, the annotations are automatically updated and saved to disk when the user visualizes the next annotation. This allows the user to rapidly skim through the annotations and adjust them without the need of going through the whole file. Furthermore it eliminates the apparently negligible overhead time for manually loading individual files, displaying them and locating interesting segments before performing the annotation. We argue that this procedure will significantly lower the time needed by an annotator to manually label segments of interest within a big speech corpus.

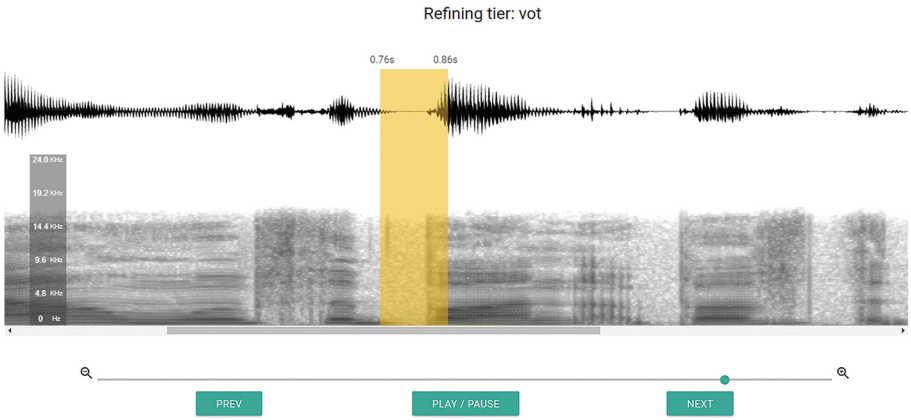
Once a set of sufficiently accurate time intervals is obtained, the user can apply a pretrained AutoVOT model on the whole dataset. This is a fully guided and customizable operation that does not require any programming skill to interact with AutoVOT. AutoVOT's parameters can be configured through VOTEUS and the output is saved directly into the textgrid of each audio file. Among the customizable parameters, the user can select a pretrained model, the dataset on which to apply it and the tier name on which to search for VOTs. All the other parameters that the

[21] Available languages are English, French, Italian, Spanish, Mandarin Chinese, Catalan, Polish, Portuguese, Southern Min, Cantonese, Japanese, Korean and Naija.

[22] Note that we refer to a generic tier present in a textgrid, which includes the output of intermediate steps of our annotation procedure. This procedure in fact will also be used at a final stage to manually refine VOT tiers provided by AutoVOT, adding a layer of human supervision to assess the quality of the predictions and correct them if needed.

original implementation of AutoVOT offers, such as the size of search window and the minimum and maximum length of the detectable VOTs, are fully controllable from the interface. This is sufficient to use VOTEUS as a proxy module to test AutoVOT, but if the final goal is to obtain accurate VOT annotations, the user can rely on the aforementioned fast tier refinement procedure to check and eventually adjust the predicted tiers. The advantages of this semi-automatic annotation pipeline therefore reflect on two important use cases: testing AutoVOT to obtain VOT predictions and precisely annotating a corpus with a head start provided by AutoVOT's predictions.

Figure 4 - *Users can refine tiers in sequence, rapidly skimming through the whole dataset. For precise refinements the annotation can be zoomed in and out within the interface*

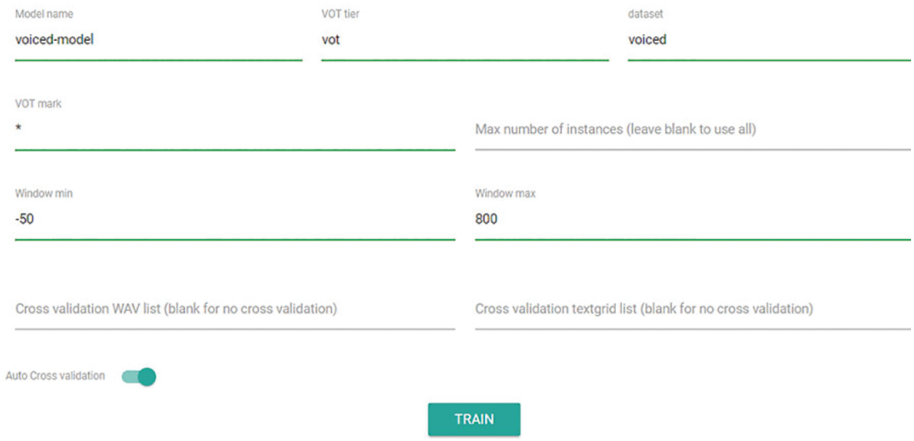


2.4 Training AutoVOT models

To obtain VOT detections it is necessary to use a functional AutoVOT model. Whereas pre-trained models can be downloaded along with AutoVOT and integrated with VOTEUS, one could need to train a model suitable for the data at hand. Through VOTEUS we permit to train new models on a custom speech corpus providing another guided procedure. Similarly to the AutoVOT evaluation functionalites, no programming is required and everything is configurable through our interface. The annotations required to train the model can be selected from an existing tier in the textgrid or manually defined by the user. The user can customize the training procedure setting all the parameters expected by AutoVOT. In Figure 5 the training interface is shown. The required parameters that the user has to set are a name to save the model, the dataset to use for training and the name of the tier with the VOT annotations. In addition there are optional parameters for AutoVOT such as the VOT mark to select a subset of annotations (e.g. "vot", "pos", "neg"), the number of instances to be used and the left and right boundaries of the annotation window in milliseconds relative to the VOT interval. The user can also decide whether to perform cross validation during training and if so which files to use as

validation set. The files for cross validation can be explicitly listed or selected by random through the selection of the "auto cross validation" option. The generated model is then saved into VOTEUS in order to be tested on new data. Again, for the sake of simplicity and compatibility, we store the trained models in the same data format originally used by AutoVOT.

Figure 5 - *Users can customize all the parameters required by AutoVOT and train a model on a selected dataset*



2.5 Exporting results

All the results produced within VOTEUS can be exported and reused with external tools. We offer a choice between different data formats to export the annotations generated with VOTEUS. Users can directly save from the interface the textgrid files to be inspected with Praat and at the same time can convert the annotations in textual form as a CSV (Comma Sparated Values) or save them as *.xls files for compatibility with Microsoft Office Excel and Apache OpenOffice Calc. We believe that this will allow more flexibility for researchers, without forcing them to use a specific tool.

3. *Conclusions*

At the age of 50, it is time for Voice Onset Time to enter the field of big corpora analyses. The access to larger linguistic datasets allows researchers to ground their understanding of this sub-segmental feature on more quantitatively realistic observations; to date, this approach has proven to benefit not only acoustic phonetics and sociophonetics, but also cognitive laboratory methodologies. It is therefore necessary to abandon the traditional time-consuming research routines based on manual annotations and automatize the preparation of the materials. Our contribution aims to increase the accessibility of already existing tools for phonetic analysis,

with the final goal of assisting the researcher through the processes of text-audio alignment, VOTs segmentation and durations extraction. VOTEUS is currently in development for Linux and Windows operating systems. Future work will focus on providing quantitative estimates about the time saved using our interface, as well as the results of usability tests. A preview version of VOTEUS will be available in early 2019 at the following link https://github.com/fedebecat/VOTEUS.

Bibliography

Abramson, A.S., Whalen, D.H. (2017). Voice Onset Time (VOT) at 50: Theoretical and Practical Issues in Measuring Voicing Distinctions. In *Journal of Phonetics*, 63, 75-86.

Bane, M., Graff, P. & Sonderegger, M. (2010). Longitudinal Phonetic Variation in a Closed System. In *Proceedings of the Annual Meeting of the Chicago Linguistics Society*, 46, 43-58.

Bang, H.-Y. (2017). The Structure of Multiple Cues to Stop Categorization and Its Implications for Sound Change. Ph.D. Dissertation, McGill University.

Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M. & Yoon, T.-J. (2015). The Effect of Word Frequency on the Time Course of Tonogenesis in Seoul Korean. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M. & Yoon, T.-J. (2018). The Emergence, Progress, and Impact of Sound Change in Progress in Seoul Korean: Implications for Mechanisms of Tonogenesis. In *Journal of Phonetics*, 66, 120-144.

Berez-Kroeker, A., Gawne, L., Kung, S., Kelly, B.F., Heston, T., Holton, G., Pulsifer, P., Beaver, D.I., Chelliah, S., Dubinsky, S., Meier, R.P., Thieberger, N., Rice, K. & Woodbury, A.C. (2018). Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in our Field. In *Linguistics*, 56(1), 1-18.

Bigi, B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. In *the Phonetician*, 111-112, 54-69.

Bigi, B., Meunier, C. (2018). Automatic Speech Segmentation of Spontaneous Speech. In *Revista de Estudos da Linguagem. International Thematic Issue: Speech Segmentation*, 26(4), 1489-1530.

Boersma, P. (2001). Praat, a System for Doing Phonetics by Computer. In *Glot International*, 5, 341-345.

Chen D., Kozbur, D. & Yu, A. (2015). Pandering vs. Persuasion? Phonemic Accommodation in the U.S. Supreme Court. Working Paper.

Chen X., Xiong Z. & Hu J. (2018). The Trajectory of Voice Onset Time with Vocal Aging. In *Proceedings of INTERSPEECH 2018*, 1556-1560.

Cheng, A. (2017). VOT Merger and f0 Contrast in Heritage Korean in California. In *UC Berkeley PhonLab Annual Report*, 13(1), 281-311.

Cho, T., Docherty, G. & Whalen, D.H. (Eds.) (2018). Special Issue: Marking 50 Years of Research on Voice Onset Time. In *Journal of Phonetics*, 71.

Chodroff, E., Godfrey, J., Khudanpur, S. & Wilson, C. (2015). Structured Variability in Acoustic Realization: A Corpus Study of Voice Onset Time in American English Stops. In *Proceedings of the 18th International Congress of Phonetic Sciences.* Paper 632.

Chodroff, E., Wilson, C. (2017). Structure in Talker-specific Phonetic Realization: Covariation of Stop Consonant VOT in American English. In *Journal of Phonetics*, 61, 30-47.

Chodroff, E., Wilson, C. (2018). Predictability of Stop Consonant Phonetics Across Talkers: Between-category and Within-category Dependencies among Cues for Place and Voice. In *Linguistics Vanguard*, 4(2), 1-11.

Dijkstra, E.W. (1982). On the Role of Scientific Thought. In *Selected Writings on Computing: a Personal Perspective*. New York: Springer, 60-66.

Foulkes, P. (2015). Sociophonetics. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper 1051.

Goldrick, M., Keshet, J., Gustafson, E., Heller, J. & Needle, J. (2016). Automatic Analysis of Slips of the Tongue: Insights into the Cognitive Architecture of Speech Production. In *Cognition*, 149, 31-39.

Heeringa, W., Van de Velde, H. (2017). Visible Vowels: A Tool for the Visualization of Vowel Variation. In *Proceedings of INTERSPEECH 2017*, 4034-4035.

Henry, K., Sonderegger, M. & Keshet, J. (2012). Automatic Measurement of Positive and Negative Voice Onset Time. In *Proceedings of INTERSPEECH 2012*, 871-874.

Keshet, J., Sonderegger, M. & Knowles, T. (2014). AutoVOT: A Tool for Automatic Measurement of Voice Onset Time Using Discriminative Structured Prediction. https://github.com/mlml/autoVOT/

Lisker, L., Abramson, A.S. (1964). A Cross-language Study of Voicing in Initial Stops: Acoustical Measurements. In *Word*, 20(3), 527-565.

McAuliffe, M., Stengel-Eskin, E., Socolof, M. & Sonderegger, M. (2017). Polyglot and Speech Corpus Tools: A System for Representing, Integrating, and Querying Speech Corpora. In *Proceedings of INTERSPEECH 2017*, 3887-3891.

Reddy, S., Stanford, J. (2015). A Web Application for Automated Dialect Analysis. In *Proceedings of NAACL-HLT 2015*, 71-75.

Schertz, J., Kang, Y. & Han, S. (2017). Cross-language Correspondences in the Face of Change: Phonetic Independence Versus Convergence in Two Korean-Mandarin Bilingual Communities. In *International Journal of Bilingualism*, 23(1), 157-199.

Sheena, Y., Hejná, M., Adi., Y. & Keshet., J. (2017). Automatic Measurement of Pre-aspiration. In *Proceedings of INTERSPEECH 2017,* 1049-1053.

Singh, R., Keshet, J. & Hovy, E. (2016). Profiling Hoax Callers. In *Proceedings of the 2016 IEEE Symposium on Technologies for Homeland Security (HST)*, 1-6.

Singh, R., Keshet, J., Gencaga, D. & Raj B. (2016). The Relationship of Voice Onset Time and Voice Offset Time to Physical Age. In *Proceedings of the 41 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5390-5394.

Singh, R., Raj, B. & Gencaga, D. (2016). Forensic Anthropometry from Voice: An Articulatory-phonetic Approach. In *39th International Convention on Information and Communication Technology Electronics and Microelectronics (MIPRO)*, 1375-1380.

Solanki, V.J. (2017). Brains in Dialogue: Investigating Accommodation in Live Conversational Speech for Both Speech and EEG data. Ph.D. Dissertation. University of Glasgow.

Sonderegger, M. (2012). Phonetic and Phonological Dynamics on Reality Television. Ph.D. Dissertation, University of Chicago.

Sonderegger, M. (2015). Trajectories of Voice Onset Time in Spontaneous Speech on Reality TV. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper 903.

Sonderegger, M., Bane, M. & Graff, P. (2017). The Medium-term Dynamics of Accents on Reality Television. In *Language*, 93(3), 598-640.

Sonderegger, M., Keshet, J. (2010). Automatic Discriminative Measurement of Voice Onset Time. In *Proceedings of INTERSPEECH 2010*, 2242-2245.

Sonderegger, M., Keshet, J. (2012). Automatic Measurement of Voice Onset Time Using Discriminative Structured Predictions. In *The Journal of the Acoustical Society of America*, 132(6), 3965-3979.

Stuart-Smith, J., Rathcke, T., Sonderegger, M. & Macdonald, R. (2015). A Real-time Study of Plosives in Glaswegian Using an Automatic Measurement Algorithm: Change or Age-grading?. In Torgersen, E., Harstad, B., Maehlum, B. & Royneland, U. (Eds.) *Language Variation: European Perspectives V: Selected Papers from the 7th International Conference on Language Variation in Europe (ICLaVE 7).* Amsterdam: John Benjamins, 225-237.

Stuart-Smith, J., Sonderegger, M., Rathcke, T. & Macdonald, R. (2015). The Private Life of Stops: VOT in a Real-time Corpus of Spontaneous Glaswegian. In *Laboratory Phonology*, 6, 505-549.

Turnbull, R. (2015). Assessing the Listener-oriented Account of Predictability-based Phonetic Reduction. Ph.D. Dissertation. Ohio State University.