ToneSwiper: Facilitating manual ToDI-annotation of Dutch prosody

Matthijs Westera, Ariëlle Reitsema
Leiden University Centre for Linguistics
{m.westera, a.reitsema}@hum.leidenuniv.nl

Abstract

Manual transcription of intonation by experts remains an essential part of research on the structure and meaning of intonation across languages, as well as for developing computational methods for automatic intonation transcription. We present ToneSwiper, a Python program with a graphical user interface that facilitates manual intonation transcription in the ToDI framework (Transcription of Dutch Intonation; Gussenhoven, 2005), with possible adaptation to similar (e.g., ToBI-like) frameworks for other languages. For the trained annotator, it enables efficient ToDI transcription of speech by integrating an audio-player, a spectrogram and pitch contour plot, auto-scroll, dynamic audio stretching, and an intuitive hotkey interface that maps key sequences to ToDI elements, e.g., pressing up-down for a high-to-low accent (H*L). In this way, transcription is conducted by 'swiping' over the arrow keys on the keyboard. We present the program and its motivation, as well as a small-scale pilot study on annotation efficiency and inter-rater agreement, using a highly challenging sample of task-oriented dialogue from the Dutch Map Task Corpus (Ladd and Schepman, 2003).

Keywords: Intonation, transcription, ToDI, software, map task

1. Introduction

Intonation is a key communicative channel. It is, in part, linguistically structured, comprised of discrete, categorical events such as high and low intonation phrase boundaries, and rising and falling accents. Speakers use intonation to comment on the pragmatic status of their utterance, for instance, to clarify how it relates to the conversational goals and to the beliefs of speaker and hearer (e.g. Brazil et al., 1980; Gussenhoven, 1984; Pierrehumbert and Hirschberg, 1990; and much subsequent work; for a recent overview see Westera et al., 2020).

On the empirical side, research on intonation, like most subfields in the study of language, requires a combination of direct observation, native speaker informants, controlled lab experiments, and corpus research. The fourth of these pillars is lagging behind, with many languages – including closely studied ones – still lacking sizable and diverse corpora with intonation transcriptions (e.g., Dutch). This can in part be linked to a lack of dedicated, user-friendly tools for efficient manual transcription, and causes, in turn, an inability to develop and evaluate computational tools for automatic intonation transcription. To aid in closing this gap, we present ToneSwiper, a computer application that enables efficient manual (expert) transcription of intonation.

ToneSwiper is an open-source Python program with a graphical user interface, offering a spectrogram with a pitch track, sound playback functionalities (play/pause, rewind, audio stretching, etc.), and a largely keyboard-controlled transcription panel. In contrast to existing, more general-purpose audio analysis and transcription tools such as Praat (Boersma and Weenink, 2025) and ELAN (2025), ToneSwiper is designed exclusively for intonation transcription. This specialization enables an efficient, keyboard-centered interface with dedicated hotkeys for the different intonational categories. For instance, pressing the arrow keys `up`, `down` results in a 'falling accent', and pressing `up` with `right` results in a 'high phrase-final boundary', with additional hotkeys for manipulations such as 'delay' and 'downstep' (see Section 2). Hotkeys can be entered simultaneously with the audio playing, facilitated by on-the-fly audio stretching (i.e., slower playback without affecting pitch). This avoids the many mouse cursor actions inevitable in more general-purpose transcription tools, removing interface friction (and potential causes of repetitive strain injury) to let transcribers fully focus on the task at hand.

ToneSwiper is designed for the transcription of the aforementioned linguistically structured part of intonation, i.e., discrete, categorical events such as high vs. low accents and boundaries. This excludes *paralinguistic* aspects of intonation, comprising gradient adjustments of pitch contour and register. Specifically, ToneSwiper is currently set up for use with the ToDI-system – Transcription of Dutch Intonation (Gussenhoven, 2005) – implementing hotkey combinations for exactly the types of accents and boundaries assumed in ToDI. This hotkey mapping can easily be adapted to other intonational inventories. Note that ToneSwiper is aimed at enabling efficient, user-friendly transcription to maximize coverage, not millisecond precision – although $\leq$100ms precision is feasible (and a 'zoom-in' function that would enhance precision may be added in the future). The resulting transcriptions can straightforwardly be imported into the Praat program as a 'point tier', in which one

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5856–5863
11–16 May 2026.

H*, L*	level high/low accent
H*L, L*H	falling/rising accent
H*LH	fall-rising accent
L*HL	'delayed' falling accent
!H*, !H*L, L*!HL	'downstepped' versions
H%, L%, %	high/low/level IP-final boundary
%H, %L	high/low IP-initial boundary
%HL	non-accent initial falling pitch
H*!H	vocative chant

Table 1: Categories in the ToDI system.

could fine-tune the position of annotations or otherwise modify transcriptions if necessary, and combine them with other transcription tiers in a Praat TextGrid. ToneSwiper is open-source and cross-platform, and is available on the Python Package Index (currently at version 0.3.3), thus installable with `pip` and similar tools in the Python ecosystem.

This paper motivates and describes the ToneSwiper program, and evaluates its use for the transcription, by two annotators, of a 4-minute spoken dialogue from the Dutch Map Task Corpus (Ladd and Schepman, 2003), a particularly challenging genre characterized by rapid turn-taking, interruptions, overlapping speech and backchannels (studied in depth, using this corpus, by Caspers, e.g., 2003).

2. Background

2.1. Systems for transcribing intonation

Languages differ in the intonational contrasts they draw. Pierrehumbert (1980) aimed at formulating a phonological grammar to account for the contrastive intonation forms of English. Silverman et al. (1992) developed the influential ToBI system (Tones and Break Indices) for transcribing English intonation, which spurred much subsequent work to develop broadly ToBI-like systems for various languages (e.g., Japanese, Venditti, 2005, French, Delais-Roussarie et al., 2015; Spanish, Beckman et al., 2002). Against this backdrop, as well as in response to prior, more trajectory-based transcription systems for Dutch intonation, the ToDI system (Transcription of Dutch Intonation) was developed (Gussenhoven, 2005).

ToDI distinguishes two tones (H=high vs. L=low), and two types of events (locations) that may carry tones, namely pitch accents (*) and intonation phrase (IP) boundaries (%). Altogether, ToDI distinguishes the accent and boundary types given in Table 1. For more descriptive characterizations and clear examples of each of these categories, we refer to the ToDI website (Gussenhoven et al.).

2.2. Current audio transcription software

Two general-purpose audio analysis and transcription tools are Praat (Boersma and Weenink, 2025) and ELAN (2025, also for video). The workflow for transcribing prosody in these programs would be to load an audio file, visualize its spectrogram and an automatically extracted pitch track to aid identification of intonational events, create an annotation tier, and play-pause-and-seek one's way incrementally through the audio while adding labels to the tier, in our case the ToDI categories. One would add labels by clicking the mouse in the right location and entering the label's text content on the keyboard. Transcription of intonation in the style of ToBI for American English is reported to take between 100× and 200× the length of the audio transcribed (Syrdal et al., 2001), though note that this includes transcription of 'break indices' absent from ToDI. Our own experience with transcribing in the slightly simpler ToDI system in Praat places our own estimates around at least 60×.

Because these programs are general-purpose, they are not particularly ergonomic for intonation transcription. Foremost, given the relatively small inventory of categories in a system such as ToDI, there should be virtually no need for manual typing, as only a few intuitively oriented keyboard keys and their combinations could already cover the full range (cf. a mouse-operated menu with the available labels in Syrdal et al. (2001)). Second, these keyboard commands could in principle be pressed in sync with the audio playback to determine the resulting labels' positions, avoiding the need for most mouse usage, which is generally less efficient and potentially straining. Third, a dedicated intonation transcription tool should enable audio stretching, i.e., 'on the fly' slowing down the playback when required (and speeding it up again afterwards if desired) *without* thereby affecting the pitch; neither Praat nor ELAN currently supports this. The ToneSwiper program presented in this paper is dedicated to intonation transcription (and nothing else), and will be centered around precisely these three features.

An alternative method for accelerating transcription is to integrate a manual transcription program with automatic methods, e.g., for prior selection of likely transcription locations or for constraining the possible transcription labels based on the context while leaving the final decision to the human. For a detailed overview of such 'prompters' and other automatic method, in the context of developing automatic ToBI-like transcription for Spanish and Catalan, see Elvira-García et al., 2016. In principle, the ergonomic improvements that ToneSwiper is aiming for are orthogonal to, and compatible with, such methods.

2.3. Corpus methods in intonation research

Theories of intonation, both its phonology and its meaning and use, should be informed by corpus research. For instance, in case certain contrasts cannot reliably be detected by transcribers or observed phenomena do not fit into one of the theory's phonological categories, this could drive revision or extension of the phonological theory; conversely, adequacy for transcription can be presented in support of a theory (Pitrelli et al., 1994). The empirical success of existing theories of intonational meaning thus far remains limited, or at least difficult to assess (Westera et al., 2020), and they have primarily been applied to specific, constructed examples, leaving their applicability to naturalistic discourse doubtful. While native speakers have clear and consistent judgments about which intonation contour fits best given a broader context (e.g., a folk story in He et al. (2012)), actually explaining such preferences in terms of intonational meaning has remained an elusive goal. While the widespread adoption of corpus methods has led to important advances in the study of meaning and use of words and phrases, corpus work on intonational meaning is lagging behind.

For Dutch, to our awareness, no sizable, representative, publicly available ToDI-transcribed dataset exists. This is due at least in part to the difficulty of, and labor investment required for, transcribing intonation. Numerous works have attempted to automate intonation transcription with computational methods, as in Rosenberg, 2010 (and many works cited therein) for English in the ToBI system, and its adaptation in Hu et al., 2020 for Dutch in the ToDI system. For the present case of Dutch, while the accuracies reported in Hu et al., 2020 are high (e.g., 94.6% accuracy on pitch accent detection, 75.4% on their classification), it must be noted that the data on which their models were trained and tested consisted only of single-speaker recordings of scripted, single utterances, selected from a limited inventory of phonetic profiles, and elicited with particular intonation contours as intended by the researcher. Although the authors claim it is a 'fair representation of natural speech', it lacks the comparative messiness of truly spontaneous speech, such as unscripted dialogue. To compare, Rosenberg (2010) used a genre comparable to our own (unscripted, task-oriented dialogue; see below), reporting lower scores (e.g., 73.5% accuracy for detecting pitch accents, and 69.8% for classifying them). More recently, Zhai and Hasegawa-Johnson (2023) report F1-scores of 0.82 for pitch accent detection and 0.86 for boundary detection, without undertaking classification.

In order to unlock corpus-based methods for intonation research, as well as to develop better automatic transcription methods (and to be able to meaningfully evaluate them) for a wider range of languages, a sizable body of human transcriptions will remain necessary. The ToneSwiper program presented in this paper is meant to facilitate meeting this need. To assess its potential, we conducted a pilot study, reported further below, in which we used ToneSwiper for transcribing a highly challenging, guided spontaneous dialogue from the Dutch Map Task Corpus (Ladd and Schepman, 2003).

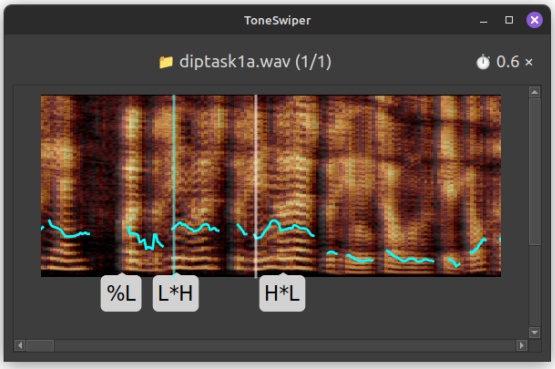

Figure 1: The main graphical user interface of ToneSwiper.

3. Description of the ToneSwiper program

3.1. User interface

After installing the toneswiper module and executing it for one or more audio files (see technical details), the graphical user interface starts. The main window automatically displays a spectrogram and pitch track for the first selected audio file, at a level of zoom adequate for visual detection and categorization of intonation events. Depending on screen resolution, between five and ten seconds of audio may fill the screen; for longer audio, the window will automatically scroll horizontally during playback to keep the visuals in sync with the audio. A vertical bar moves from left to right over the spectogram (and becomes centered as the window overflows, i.e., while scrolling) to indicate the current playback position.

Keyboard controls allow the user to add transcriptions, both while the audio is playing and while paused. Transcriptions will appear directly below the spectrogram, as selectable/editable/movable text bubbles, their horizontal position visually matching the corresponding position in the audio. Transcriptions appear immediately upon pressing the hotkeys, but at a fixed distance to the left (i.e. earlier in time) relative to the playback position: this assumes that transcription is normally 'delayed', i.e.,

H*, L*	up arrow / down arrow
H*L, L*H	up-down / down-up
H*LH	up-down-up
L*HL	down-up-down
!H*, !H*L, L*!HL	hold control
H%, L%, %	right-up, right-down, right
%H, %L	left-up, left-down
%HL	rare, no hotkey
H*!H	up-(back)slash

Table 2: ToneSwiper key combinations.

one hears a falling accent, and only e.g. half a second later one can press the appropriate hotkeys. This 'delayed' position at which transcriptions will be inserted is shown by a separate vertical bar (i.e., to the left of the playback position bar). The amount of delay can be modified during transcription, and can even be set to 0, in which case the two vertical bars coincide, and transcriptions will appear exactly at the current playback position, instead of slightly 'in the past'.

Playback speed can be decreased (and increased again) with the −/+ keys, applying real-time audio-stretching to slow down the audio without lowering the pitch. Slowing down the audio (as well as switching back to normal speed for context) is typically necessary for reliable transcription. In addition, the program offers keyboard controls for play/pause, for seek (forward and backward), for jumping to the start and end positions and to the position of the last transcription, as well as for moving between sound files. While keyboard-centered by design, transcriptions (text bubbles) can also be added, moved around, edited and deleted with mouse clicks. Pressing `F1` opens a secondary window listing the keyboard and mouse controls.

The keyboard combinations for the ToDI categories are listed in Table 2. Each annotation can freely be modified by typing in a selected text bubble, e.g., for transcriptions (including comments) not covered by the hotkeys. The key combinations are currently fixed in the program's source code, but this will be made more easily adjustable to other ToDI- (and ToBI-)like systems in the future.

3.2. Technical details

ToneSwiper is available as a Python package on the Python Package Index (PyPI), hence installable with standard programs such as `pip`. It works on Unix-like platforms and on Windows, with Python 3.10 and higher. ToneSwiper currently only supports `.wav` files; conversion to and from `.wav` is straightforward with existing tools. Running the program to transcribe one or more audio files will result in a list of pairs for each file, each pair composed of a time stamp (within the audio file) with the transcribed ToDI label (e.g., H*L, H%, etc.). The program can import and export transcriptions from/to a Praat TextGrid ('Point' tier), as well as importing/exporting transcriptions in JSON format.

Although ToneSwiper has a graphical user interface for the transcription itself, it is intentionally minimalist and offers no graphical way of choosing the `.wav` files to be transcribed. Instead, in keeping somewhat with the Unix philosophy and the advantages of a primarily text-based interface, this is done on the command-line. For instance, the command `toneswiper dialogues/*.wav --textgrid todi` will start the application for transcribing all `.wav` files in the (hypothetical) `dialogues` folder, and will load the 'todi' tier from (and save to) analogously named `.TextGrid` files if they exist – otherwise it will create such files.

The ToneSwiper source code is publicly available on GitHub. It relies on PyQt6 for its graphical user interface, on Parselmouth (Jadoul et al., 2018) for Python bindings for the Praat program (Boersma and Weenink, 2025) for extracting the spectrogram and pitch track, Matplotlib for visualising it (Hunter, 2007), the package 'tgt' (TextGrid tools; Buschmeier and Wlodarczak, 2013) for reading and exporting Praat's TextGrid format, on Soundfile (Bechtold) and PyLibRb for audio stretching (Głomski; Python-bindings into the Rubber Band Library), and besides that Numpy (Harris et al., 2020), Sounddevice and the Python standard library. Toneswiper uses semantic versioning and is provided under the European Union Public License (EUPL1.2) for free/open source software.

4. Pilot study

4.1. Method

In order to evaluate the usability of the ToneSwiper tool, especially on naturalistic dialogue, we doubly annotated a 4 minutes and 15 seconds long dialogue from the Dutch Map Task Corpus (Ladd and Schepman, 2003). The full corpus consists of 8 task-oriented dialogues, following the set-up of the original HCRC Map Task Corpus (Anderson et al., 1991). Each conversation revolved around the goal of reproducing a route on a map via (only) verbal collaboration. Both participants have a map in front of them, which for one participant – the instruction giver – includes a printed route, to be explained verbally to the instruction follower, who needs to draw that same route on their map. Crucially, small discrepancies between the reference points on the two maps complicate the conversation, yielding more interactive turn-taking dynamics and more varied intonation contours.

The transcribers, the authors of this paper, are both linguists with prior training in listening for ToDI

categories, and with some prior experience in conducting ToDI-transcription using the existing Praat software. Neither of us had listened to this specific dialogue before. The dialogue was cut into five fragments of around 50 seconds each. After transcribing the first fragment, we met to align our annotations and discuss any transcription differences. We then annotated the remaining four fragments.

Dialogues in the map task genre are spontaneous (albeit task-driven), and involve rapid turn shifts, interruptions, overlapping speech, backchannels ('okay…'), filled pauses ('ehm…') and laughter, making it challenging to transcribe. The dialogue we selected has 800 words from both speakers combined (including, e.g., backchannels), for an average rate of 190 words per minute. The instruction-giver and instruction-follower were likely not recorded on separate microphones, as our copy of the corpus has only a single audio channel for both speakers. Fortunately, it was easy enough to tell apart the voices of the two speakers, who were male and female in this dialogue. We transcribed one speaker at a time, such that each audio fragment was transcribed twice. This was easier than transcribing both speakers at once, both in terms of cognitive load and for saving the transcriptions of each speaker in a separate transcription tier (essential for overlapping speech). Each of us transcribed the different audio fragments and the two speakers in each in the same order.

While we will report transcription rate and error rate, it must be kept in mind that this is meant only as a pilot study. For one, transcription speed and agreement rates are highly dependent on the audio being transcribed, and the scope of this pilot is limited to one specific dialogue between two particular speakers. Moreover, as we extend our transcription efforts in the future we will continue to align our transcriptions (as we currently did only for the first fragment), and will converge on better and more consistent transcription choices. As such, we will report quantitative findings with no greater intended weight than our qualitative experiences.

4.2. Transcription rate

Transcribing the 4 minutes and 15 seconds long dialogue took on average 170 minutes per transcriber, or 40 minutes to transcribe one minute of audio. This resulted in an average of 738 transcribed ToDI events (accents, boundaries) per transcriber for the full dialogue, or 2.9 transcribed events per second of audio (on average per transcriber); almost one event per word. This high event density is in part a consequence of the many short utterances such as backchannels, typically a single word ('okay…') but often three discernible ToDI events (initial boundary, pitch accent, final boundary). Put differently, we spent on average 10 seconds per transcribed ToDI event. This crucially includes listening to the entire dialogue (mostly at decreased playback rate), as well as frequent rewinding and re-listening as required.

These rates show that intonation transcription remains a challenging, cognitively demanding task, especially for naturalistic dialogue. The ToneSwiper program can accelerate this only to the extent that it removes interface friction. In our experience as transcribers, it indeed removes such friction to a large extent, compared to our experience transcribing intonation in general-purpose programs such as Praat. Note that, within the scope of this pilot, neither of us reached sufficient automatism to immediately translate the events we heard into the right key combinations to press, so some gains in transcription rate are expected to come with increased familiarity.

4.3. Inter-rater agreement

As mentioned, after transcribing the first audio fragment, we manually aligned our labels to identify and resolve any systematic differences, before continuing with the remainder. Our qualitative impression from this manual alignment of the first audio fragment is that many transcription differences are resolvable, namely, that upon attentive relistening there is typically a favored analysis. After this, we transcribed the remaining four audio fragments. In the following analysis we include all five fragments, and without adjusting any transcriptions based on our manual alignment. This is because the purpose here is to investigate the result of two transcribers independently using the tool.

To assess inter-rater agreement, we need to determine which of our respective annotations concerned the same perceived intonational event, and which did not. For instance, does a high accent (H*) annotated by one transcriber belong with the other transcriber's falling accent (H*L) 80 ms earlier (say), or with their high boundary tone (H%) 110 ms later? We automatically determined a plausible alignment by using a 'greedy' algorithm that takes into account the a priori likelihood of certain errors, on the basis of our manual alignment for the first audio fragment. The algorithm iteratively identifies 'the next best match' of transcription labels to pair, while gradually relaxing its criteria for doing so. Initially, the 'best match' for a given transcription is the nearest (in time) transcription (by the other transcriber) within 100 ms that has the exact same label (and which has not yet been consumed by a better match). This criterion is then relaxed to permit matching transcriptions, still within 100 ms, that differ slightly (for instance H*L with H*), followed by string-identical transcriptions within 200 ms, then slight mismatches within 200ms, then more severe mismatches, and so on, up to quite severe mis-

matches as long as both labels are still available and within at most 300ms of each other. We did not extensively tweak the parameters and thresholds of this approach, but settled on what appeared by manual inspection of the resulting alignment to be a good solution. (Automatically tuning the parameters of this algorithm would require more data.)

The greedy alignment process identified 609 places where both transcribers annotated an event (this can be compared and contrasted with the, on average, 738 events per transcriber, or to the 866 places where at least one transcriber entered something), and these were on average 82ms apart (standard deviation = 118). Of these, 376 were given identical ToDI labels by both transcribers, and these were on average 60ms apart (standard deviation = 54), i.e., in cases where both transcribers agreed on the ToDI label, their transcriptions fell within, on average, 60ms of each other.

The confusion table in Figure 2 shows patterns of agreement and disagreement (where 'disagreement' means the transcribers transcribed a given event differently). The top row and left column represent cases where only one of the annotators transcribed an event – the top-left cell being empty because there are uninformatively many places where both transcribers did not annotate anything. In 73% of the cases where at least one transcriber annotated an event, the other transcriber did so too (percentage not shown in the image, but computable from it); in the remainder (27% of cases), one transcriber transcribed an event while the other did not (or there was a candidate match, but it had already been paired to a better match by the greedy algorithm). Setting aside the first row and column of the confusion table, the diagonal, representing alignment, is the most populated, and misalignments tend to be clustered (visible squares): first a cluster concerning initial boundary tones, then one concerning the final boundary tones, and then a cluster representing confusions between the various accent types.

The first cluster of disagreements shows considerable confusion (in the technical sense) between low and high initial boundaries (%L, %H), which the ToDI guide notes can be difficult to distinguish (especially %L immediately followed by a high accent); though the transcribers were in agreement in around two-thirds of the initial boundaries. Notably, there is no clear asymmetry between the two transcribers; both regularly transcribe %L where the other transcribed %H, and vice versa. Likewise, both transcribers (if we look at the top row, and first column) regularly miss low or high boundaries where the other does transcribe them. This suggests a lack of clarity for both transcribers, and/or a genuine impossibility to reliably identify the presence of the underlying phenomenon.

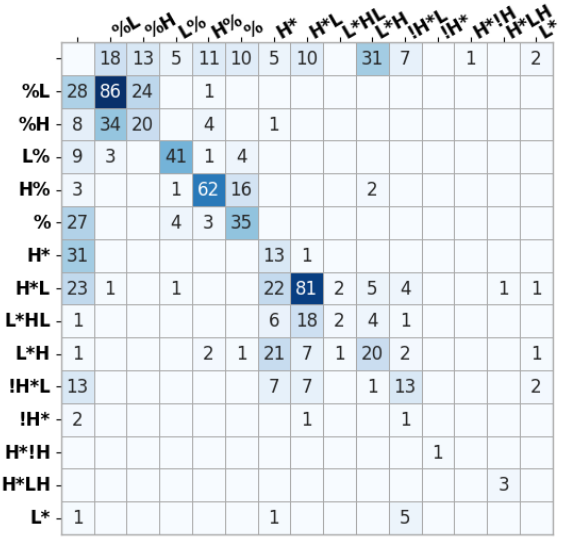

Figure 2: Confusion table of the two transcribers.

The next cluster of disagreements pertains to the final boundary tones, revealing decent agreement but with some confusion between high and level boundaries (H% vs. %). These can be difficult to distinguish at times, especially on short utterances with a rising accent (L*H). This time, an asymmetry is visible, with one transcriber more regularly perceiving as level boundaries ones that the other transcribed as high; at the same time, they regularly transcribed a level boundary where the other transcribed no event at all (i.e., the number 27 in the left-most column).

The third cluster of disagreements reveals some difficulty distinguishing the different accent types, with one transcriber hardly transcribing rising accents (L*H), and the other transcribing fewer falling accents (H*L). Indeed, one transcriber annotates many rising accents where the other identifies none (L*H; number 31 in the top row), for which high boundary tones after a low pitch (e.g., preceded by a falling accent) may be to blame, as these can result in prominent rises. One transcriber more readily transcribed 'delayed' falling accents (turning H*L into L*HL, audible as a rise toward the high tone), transcribing 18 such cases where the other transcribed a plain, non-delayed fall (central in the plot, below the number 81). A similar difference may underlie the 28 rising accents (L*H) identified by one transcriber, which the other transcribed as high (H*, 21 cases) or falling accents (H*L, 8 cases). Downstep (!) on high or falling accents is detected fairly reliably by both, though one transcriber occasionally opts for a low accent (L*) instead.

Altogether, transcription of intonation remains a challenging task, certainly on naturalistic dialogue. Fortunately, there are some clear patterns in the inter-rater disagreements, and manual alignment

and discussion of the first audio fragment pointed to the majority of disagreements being resolvable upon closer inspection. Therefore, additional manual alignment and discussion sessions could increase familiarity with some of the remaining pitfalls, and thereby increase inter-rater agreement. Indeed, our findings might also hint at the necessity of frequent inter-rater discussions as an integral part of multi-rater transcription. Such a workflow, too, would be facilitated by an efficient initial transcription round, after which only the misalignments would require further attention.

5. Conclusion and outlook

This paper motivated and presented ToneSwiper, an open-source program for manual, expert transcription of intonation according to the ToDI system. Our small pilot, in which two transcribers used ToneSwiper on naturalistic dialogue from the Dutch Map Task Corpus, confirmed its added value as a transcription tool. Its keyboard-centered design, on-the-fly audio-stretching and auto-scroll make for a pleasant transcription workflow, by removing the main sources of interface friction present in more general-purpose transcription programs.

ToneSwiper can accelerate intonation transcription only to the extent that it removes interface friction (and associated cognitive load); it does not reduce the difficulty of the task itself, i.e., that of detecting ToDI events by listening to audio and viewing the spectrogram and pitch track. In the case of our pilot, we chose the particularly challenging genre of naturalistic dialogue. This resulted in average transcription rates of around 40 minutes per minute of audio. This compares favorably to the 100×-200× rates reported in Syrdal et al., 2001 for ToBI (though including break indices) and to own estimate of at least 60× for transcribing ToDI in Praat. Efficiency gains by using ToneSwiper will be greater for more clearly spoken, single-speaker genres, such as broadcast news or audiobooks.

One remaining source of friction in the current version of ToneSwiper was the 'rigidity' of the transcription locus at a fixed (though customizable) delay behind the audio playback position. This meant that pausing and rewinding were frequently necessary as the opportunity to transcribe at a specific location could be easily missed. To remove this remaining hurdle, in future versions of ToneSwiper we plan to implement the use of cursor/touchscreen swipes anywhere on the spectrogram during playback, in order to be able to flexibly transcribe any position in the current window (as opposed to only at a fixed delay behind the playback position).

While the aforementioned delay (combined with audio stretching) was designed to allow 'on the fly' transcription, during this limited pilot neither transcriber reached the level of automatism required for translating detected intonational events into the right key combinations right away, synchronously with the (slowed-down) audio playback. Deciding on the transcription and finding the corresponding hotkeys proved challenging to combine with continued attentive listening for subsequent events as the audio kept playing. More experience with the tool is expected to improve this, although the current pilot study is too small, and especially too varied in terms of event density of the different audio fragments, to quantitatively demonstrate the onset of such an effect. Genuine 'on the fly' (albeit slowed-down) transcription is expected to be more feasible with less challenging audio genres.

6. Acknowledgements

This publication is part of the project *Who's next? The role of speech melody in the turn-taking system of Dutch* with file number 406.22.CTW.004 of the research programme SSH Open Competition M 2022, which is (partly) financed by the Dutch Research Council (NWO).

7. Bibliographical References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Bastian Bechtold. *Soundfile [Python package]*.

Mary E. Beckman, Manuel Díaz-Campos, Julia Tevis McGory, and Terrell A. Morgan. 2002. Intonation across Spanish, in the Tones and Break Indices framework.

Paul Boersma and David Weenink. 2025. *Praat: doing phonetics by computer*.

David Brazil, Malcolm Coulthard, and Catherine Johns. 1980. *Discourse intonation and language teaching*. Longman, London.

Hendrik Buschmeier and Marcin Wlodarczak. 2013. TextGridTools: A TextGrid processing and analysis toolkit for Python. In *Tagungsband der 24. Konferenz zur elektronischen sprachsignalverarbeitung (ESSV 2013)*.

Johanneke Caspers. 2003. Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, 31(2):251–276.

Elisabeth Delais-Roussarie, Brechtje Post, Mathieu Avanzi, Carolin Buthke, Albert Di Cristo, Ingo Feldhausen, Sun-Ah Jun, Philippe Martin, Trudel Meisenburg, Annie Rialland, et al. 2015. Intonational phonology of French: Developing a ToBI system for French. *Intonation in romance*, pages 63–100.

ELAN. 2025. *Version 7.0 [Computer software]*. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen.

Wendy Elvira-García, Paolo Roseano, Ana María Fernández-Planas, and Eugenio Martínez-Celdrán. 2016. A tool for automatic transcription of intonation: Eti_ToBI a ToBI transcriber for Spanish and Catalan. *Language Resources and Evaluation*, 50(4):767–792.

Carlos Gussenhoven. 1984. *On the grammar and semantics of sentence accents*, volume 16. De Gruyter Mouton, Berlin/New York.

Carlos Gussenhoven. 2005. Transcription of Dutch intonation. *Prosodic typology: The phonology of intonation and phrasing*, 118:145.

Carlos Gussenhoven, Toni Rietveld, and Joop Kerkhoff. Transcription of Dutch Intonation, Second Edition, version 2.3.

Paweł Głomski. PyLibRB [python package].

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Xuliang He, Vincent J. van Heuven, and Carlos Gussenhoven. 2012. The selection of intonation contours by Chinese L2 speakers of Dutch: Orthographic closure vs. prosodic knowledge. *Second Language Research*, 28(3):283–318.

Na Hu, Berit Janssen, Jeff Hansen, Carlos Gussenhoven, and Aoju Chen. 2020. Automatic analysis of speech prosody in Dutch. In *Interspeech*, pages 155–159.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

D. R. Ladd and A. Schepman. 2003. Dutch map task corpus, 1999. SN: 4632.

Janet Pierrehumbert. 1980. *The phonology and phonetics of English intonation*. Thesis, Massachusetts Institute of Technology. Accepted: 2009-01-23T14:36:47Z.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*, pages 271–311. MIT Press, Cambridge, MA.

John Pitrelli, Mary Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling. In *Proceeedings of the 3rd International Conference on Spoken Language Processing*, pages 123–126. Yokohama, Japan.

PyQt6. *[Python package]*.

Andrew Rosenberg. 2010. AutoBI: a tool for automatic ToBI annotation. In *Interspeech*, pages 146–149. Makuhari, Chiba, Japan.

Rubber Band Library. *[Computer software]*.

Kim E.A. Silverman, Mary E. Beckman, John F. Pitrelli, Mari Ostendorf, Colin W. Wightman, Patti Price, Janet B. Pierrehumbert, and Julia Hirschberg. 1992. ToBI: a standard for labeling English prosody. In *ICSLP*, volume 2, pages 867–870.

Sounddevice. *[Python package]*.

Ann K. Syrdal, Julia Hirschberg, Julie McGory, and Mary Beckman. 2001. Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, 33(1):135–151. Speech Annotation and Corpus Tools.

Jennifer J. Venditti. 2005. The J_ToBI model of Japanese intonation. *Prosodic typology: The phonology of intonation and phrasing*, pages 172–200.

Matthijs Westera, Daniel Goodhue, and Carlos Gussenhhoven. 2020. Meanings of tones and tunes. In *The Oxford Handbook of Language Prosody*. Oxford University Press.

Wanyue Zhai and Mark Hasegawa-Johnson. 2023. Wav2ToBI: a new approach to automatic ToBI transcription. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, pages 2748–2752.