TAL2018, Sixth International Symposium on Tonal Aspects of Languages
18-20 June 2018, Berlin, Germany

ATLAS (Automated Tone Level Annotation System): A tonologist's and documentarian's toolkit

Emily Grabowski[1], *Laura McPherson*[1]

[1]Dartmouth College, United States

Emily.j.grabowski.18@dartmouth.edu, Laura.e.mcpherson@dartmouth.edu

Abstract

This paper describes a novel computational toolkit for tonal analysis: ATLAS (Automated Tone Level Annotation System). Tone remains a challenge in many language documentation projects, and far too often still, one comes across descriptive and theoretical treatments of tone languages in which tone marking is entirely absent or of questionable accuracy. ATLAS takes as its input a WAV file and TextGrid delimiting tone-bearing segments and outputs normalized pitch level annotations intermediate between raw f0 and phonemic categories. These "tone level" annotations represent a discrete numerical version of the dashes often used as a broad phonetic transcription of tone. The number of levels can be set by the researcher, and a number of raw phonetic measures are also outputted by the tool. ATLAS is designed to be used by anyone regardless of experience with tone or computational methods, thus promoting the inclusion of objective, replicable pitch data in documentary, descriptive, or theoretical materials on tone languages. We also show the utility of ATLAS's broad phonetic annotations in understanding the surface realization of already determined phonemic categories and in making hypotheses about unanalyzed tone systems.

Index Terms: tone, phonetics, pitch, computation, documentation

1. Introduction

Even after countless articles have countered the narrative of tone being exotic and have offered concrete guidance on how to analyze a tone system, tone still continues to intimidate many students and professional linguists alike. The unfortunate consequence of this attitude is that tone is often ignored or underanalyzed in language documentation and description.

When materials do contain tone marking, the transcriptions are by and large phonemic. This is of course natural and desired, but given the abstract nature of many tone systems, there can be quite a gap between the phonological annotations and what is happening at the level of f0. Especially if there is no available description of the phonetic realization of tone, phonemic annotations can be of little use in reconstructing the actual pronunciation of a word or phrase, which could enable us to reanalyze the tone system later as our theoretical frameworks change. Worse, when researchers are uncomfortable or untrained in dealing with tone, tone marking may in fact detract from rather than add to analyzing the tone system.

This paper describes a computational tool designed to address both cases: where tone would be otherwise left unmarked and where the only tonal annotations are phonemic. ATLAS (Automated Tone Level Annotation System) takes a recording as an input and outputs normalized pitch annotations intermediate between raw phonetics (f0) and a phonemic analysis. The level annotations created by ATLAS are designed to create an objective, replicable, and digitizable version of the messy system of dashes often found as a descriptive lingua franca for the realization of tone (see §2). We aim to show the utility of including such a level of annotation in documentary materials both to assist current analysis and to make the materials maximally useful for future researchers. Note that ATLAS is **not** designed to produce phonological annotations, nor do we argue that intermediate tonal representations should replace the need for phonemic analysis. Rather, they provide another transparent level of pitch data that can be produced by anyone, regardless of their experience with tone, and which can help answer questions about tone that phonemic annotation alone cannot. Beyond this immediate goal, we show that ATLAS also has a number of useful and easy-to-use functions for tone research, including pitch extraction, correction, and normalization, duration measurements, and logging information about an individual speaker's pitch across a corpus of recordings. While individual Praat or Python scripts may perform one or more of these functions, ATLAS groups them together in one tool designed to be used by even those with no coding experience.

In existing work on tone languages, we can characterize two types of annotation: broad phonetic annotation and phonological annotation. The former is typically presented in the first few pages of a description and is meant to capture the surface realization of tone in an easily digestible manner. Here, the most common descriptive *lingua franca* is either numbers or dashes arranged on a vertical axis, approximating IPA tonal characters like ˥, ˦, ˧, etc. For instance, consider the following annotation, recreated from [1] for the Oceanic language Numèè (New Caledonia), where both numbers and dashes are used:

(1) gú càpē pàṇą̄ą́ kò ɲú wìi tó
1 2 1.5 3 2.5 4 3.5 4.5 4.5

The numbers range from 1 (the highest level) to 4.5 (the lowest level in the utterance), captured by the ever-descending series of dashes. The issue with this approach is that it is not clear whether the numbers represent simply the highest pitch in the utterance or whether it is the pitch ceiling for the language as a whole; similarly for the lowest pitch level. Further, the horizontal dashes run the risk of being unsystematic, hard to interpret, and virtually impossible to digitize and search.

Of course, in most work on tone languages, phonetic annotations of this sort are abandoned as soon as a phonological analysis is in place. At this point, only phonemic categories are marked (typically through diacritics or tone numbers). While this is unquestionably a desirable aspect of language transcription (with phonemic tones being in principle representative of the speaker's cognitive categories), phonemic

10.21437/TAL.2018-16

annotations are not without problem. Notably, without a thorough description of tonal phonetics, they obscure the surface realization of tone, which may be strikingly different due to effects such as downdrift, downstep, upsweep, tonal absorption, high carryover, and so on. [2] And given the widespread lack of training in working with tone, there is no guarantee that a researcher's tonal analysis is in fact correct. If the recording is available, the user could refer back to the acoustic signal, or look at pitch tracks, but this requires some familiarity with acoustics to be interpretable, is prone to sampling errors such as doubling and halving, and varies between speakers to the extent that including raw f0 numbers may hinder a general understanding of the tone system.

We are thus faced with a dilemma regarding documentary materials of tone languages: Not marking tone leaves out a crucial part of the language's morphophonology and results in materials of little use for future phonological research. But marking only an abstract level of phonemic tone can potentially propagate a misunderstanding of the tone system, with too few or too many levels and little to no indication of surface melodies. To maximize the usability of documentary materials for other researchers and language learners and to avoid relying on the transcriber's ear, we need a tool to facilitate the inclusion of broad phonetic tonal annotations that are both objective and replicable, ideally alongside phonological annotation. Note that similar arguments have arisen in the literature on intonation, e.g. regarding ToBI annotations [3].

2. Previous tools for tonal analysis

Previous work in this area largely focuses on providing automatic phonemic categorization of tone. A common method is based on the Hidden Markov Model (HMM), which is trained on manually annotated data to identify tonal categories. HMMs have been implemented in languages such as Mandarin [4][5], Thai [6], and Cantonese [7]. Other techniques used to address this question include neural networks, which may require less training data and typically aim to increase the speed of transcription for languages with known tone systems [8].

A second technique that has been leveraged in tone analysis is clustering. For instance, [9] leveraged *k*-means clustering to identify each of Khamti's four surface tone realizations. The computational model performed well for citation tones, particularly for the three contour tones, but was less successful for tones in context. This approach does not require any training data but again focuses on identifying phonemic categories. Another type of language-independent clustering is implemented in the software Toney [10]. This tool aids the user in grouping perceptually similar tones together with the goal of faster identification of phonemic categories.

Some research has also addressed the converse: reconstruction of pitch tracks from phonemic annotations. One such example is text-to-speech in African tonal languages [11]. In this case, the goal is naturalistic speech synthesis for known tone systems (not a trivial task, as laid out §2).

While each of the above technologies address important issues in tone research, they do not fill the same need as ATLAS. ATLAS generates a set of broad phonetic annotations that are tied directly to the f0 track of an utterance. The primary purpose of ATLAS is to promote transparency and replicability in documentary materials, but the output may also support the researcher in a variety of applications (see §5 below).

3. ATLAS

3.1. ATLAS input and workflow

ATLAS is implemented in the open-source programming language Python. All parameters are set using a simple graphical user interface (Figure 1), removing the need for researchers to interact directly with Praat or Python. The source code, however, is made available in addition to the compiled tool, meaning parameters within the program (e.g. outlier criteria) can be changed to suit the user's needs. To download a beta version, visit http://www.dartmouth.edu/~mcpherson/

The tool takes as input an audio file (in WAV format) and accompanying TextGrid, or a directory containing audio and TextGrid files for a single speaker. The TextGrid is annotated to indicate target segments for analysis. Any segment that generates a pitch track may be annotated as a target for analysis. Common targets include syllable nuclei and sonorant codas.

ATLAS uses Praat's pitch-tracking algorithm to extract f0 information from the target segments by measuring the f0 as generated by Praat's algorithm every 10 ms.

A major benefit of ATLAS is that it will not only extract f0 information but also apply algorithms to automatically clean the f0 information. ATLAS addresses three kinds of errors in pitch extraction from acoustic recordings. First, the pitch extraction algorithm may fail to find any f0 for a given sample, or for several samples in a token. If the undefined sample is located on the boundary of a token, the problematic samples are removed but the rest of the token is retained for analysis. If the undefined sample is not located on a boundary, then the token is deemed too problematic to retain and is excluded from further analysis. Also, if more than 25% of the token's samples do not contain f0 information, including those on boundaries, the token is excluded.

The second error type that may occur is commonly termed doubling or halving. A syllable with an f0 of 100Hz has a subharmonic at half of that frequency (50Hz), a harmonic at twice the frequency (200Hz) and so on. Since pitch-extraction algorithms are estimating the pitch based on the frequency of the signal, in some cases they will return a harmonic or subharmonic frequency rather than the true frequency of speech. ATLAS will locate and exclude tokens with sudden jumps in the sampled f0, a sign of doubling or halving.

After these first two error types have been identified and the tokens excluded, ATLAS makes a pass to filter out remaining outliers. Any samples more than three standard deviations from the mean are excluded from further analysis.

After the data are cleaned, the f0 values undergo normalization. We follow a widely practiced normalization procedure [12][13][14][15][16], etc. and normalize f0 to semitones (for a comparison of semitones to other normalization procedures, see [17]). We choose here to normalize to a speaker-specific f0, which is the median of the speaker's range. This allows for better between-speaker comparison than using raw f0 alone. If ATLAS is given an input directory containing multiple recordings from a single speaker, the tool will take the median over all recordings. Thus, all recordings analyzed by the tool for a given speaker will be normalized to the same reference f0. This both gives a better picture of a speaker's overall range rather than the range used in a particular recording (which may be broader or narrower than typical) and ensures consistency across recordings.

Finally, ATLAS automatically assigns each segment a 'tone level', which is essentially a numerical representation of the series of dashes used as a descriptive *lingua franca* for phonetic

tone annotation. This step considers all f0 information from all recordings provided to ATLAS and uses as a maximum and minimum value the 99th and 1st percentile of the speaker's range to mitigate the effect of potential remaining outliers. All values that are more extreme than these values are automatically assigned to either the highest or lowest bin. The speaker's range after normalization is divided into equal bins, or levels, the number of which can be determined by the researcher to achieve the desired level of detail. The levels are labeled numerically such that 1 refers to the lowest level. Each sample is assigned 2-3 bins (parameter set by the researcher) to capture contours in the pitch track. Bins are assigned based on the pitch at 20% and 80% (and optionally 50%) of the way through the target to reduce consonant effects on f0.

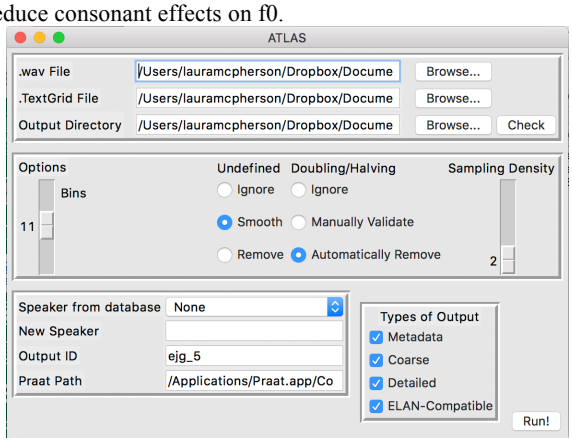

Figure 1: Graphical user interface (GUI) for ATLAS.

3.2. ATLAS outputs

The output of ATLAS is useful in a variety of applications. The tool extracts f0 from the recording using the widely used Praat algorithm, performs automated data cleaning, and normalizes the f0 for further analysis. It can also batch process large amounts of data and both normalize and discretize data across recordings using the same parameters and normalization values.

Tab-delimited outputs contain this information for every f0 measurement (every 10 ms), or at 2-3 points per segment. These outputs collate raw and normalized measurements in one place: f0, semitones, duration, and outlier detection, in addition to the broad phonetic tone levels.

In addition to the tab-delimited text outputs, ATLAS produces TextGrids containing targets and accompanying tone level annotations. These can be used either directly in Praat or integrated with an ELAN project containing other levels of annotation (phonemic transcription, interlinear glosses, syntactic category, etc.); see Figure 2 for an example. Used in conjunction with phonological annotations, tone levels illustrate the general behavior of surface realizations of pitch. It also allows for easier interpretation of the pitch track by human analysts, especially when visually processing annotations in a publication or in ELAN. For example, in Figure 2 (illustrating Seenku, a four-tone Mande language, with the following ATLAS parameters: 11 levels, two levels per target), the super-high tone of the first word is realized at the top of the speaker's range, but the following extra-low tone is only realized as a fall to the middle of the range, while the final high tone continues to rise throughout, characteristic of phrase-final intonation in the language. Unlike the old system of dashes, these annotations are easily understandable, searchable, and able to be included in more detailed annotations as illustrated here.

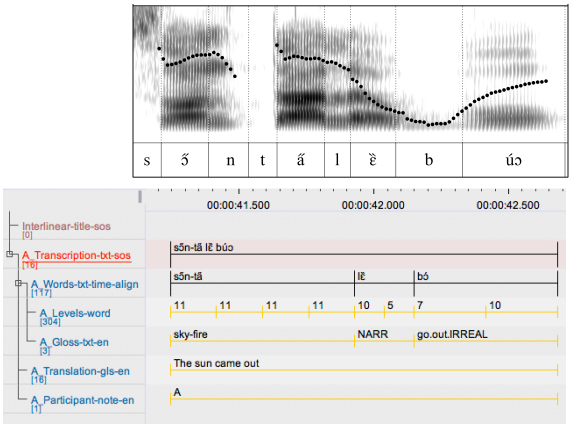

Figure 2: ATLAS levels as a tier in ELAN annotation, derived from f0 shown in the panel above.

If the transcriber would otherwise not mark tone in the transcription, ATLAS can be used to automate the inclusion of pitch data that can help others analyze the tone system.

4. ATLAS annotations and tonal analysis

4.1. Investigating the surface realization of tone categories

If a phonological analysis of the tone system is already in place for a language, ATLAS annotations can help identify phonetic, phonological, and intonational processes that affect the realization of tone categories.

We examined the ATLAS output (11 bins, 2 samples per target) for three illustrative recordings of Seenku, a four-tone Mande language spoken in Burkina Faso. Two of the recordings consisted of the same elicitation list produced by two speakers, one male and one female, where each target word was embedded in the frame sentence *ăa* ___ *sɑ̀* 's/he bought ___'. The third recording was of a different male speaker recounting the North Wind and the Sun (NWAS) translated into Seenku.

First, multiple repetitions of the frame sentence allow us to investigate whether there are any effects of context on the initial LS (low-superhigh) rising tone; this is natural to ask, since LS undergoes simplification in many environments in Seenku. For the female speaker, LS had a mean tone level pronunciation of 5.4-10.2 (coming in just shy of the top of her range), and the male speaker's mean pronunciation was 5-9.1. Surprisingly, the realization of LS does not change significantly depending on tonal context. Before an extra-low (X) tone, the female's LS rise shows the mean levels 4.7-9.8, while before super-high (S), the mean levels are 5.2-9.7. This lack of contextual effect may be due to a stronger prosodic boundary between the subject and the following object. Within a phonological phrase, on the other hand, LS is often simplified to L before S.

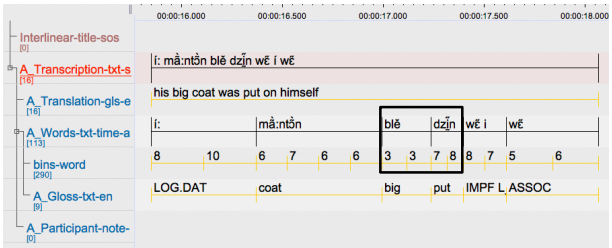

Figure 3: ATLAS annotations in ELAN for /blɛ̌ dzǐ/ → [blɛ̀ dzǐ]

As shown in the NWAS excerpt in Figure 3, the LS+S sequence is realized as L+S on the surface; the S portion of the LS rise is completely absorbed into the following S tone, which has also undergone downdrift (reaching a level of only 8 out of 11). Figure 2 above from the same story likewise demonstrated phenomena such as high carryover and intonational rises visible in the ATLAS tone level annotations.

In short, the tone level annotations created by ATLAS reveal differences in surface realization that are obscured by marking phonemic categories alone, showing the utility of including a broad phonetic level of tone transcription in materials even if the tone system is relatively well understood.

4.2. ATLAS output and clustering

If the tone system is not yet well understood, we show here that the tone level output of ATLAS can be used to help identify potential tonal categories with as much accuracy as either raw f0 or normalized semitone measurements.

We tested ATLAS's sensitivity by application to prediction the surface tonal categories of a dataset using two speakers, one male (n=106 targets) and one female (n=132 targets). We ran a *k*-means clustering algorithm (*k*=5) on the ATLAS output for each speaker using F0 measured in Hz, semitones, and three different numbers of discrete levels (5, 8, and 12).

For this analysis, we used the elicited frame sentence data introduced above, focusing only on the target words (i.e. excluding the frame sentences themselves). This produced much cleaner and less variable data. Each word was annotated by a human researcher with the phonemic tone category (four level tones, X, L, H, and S and one contour tone HX).

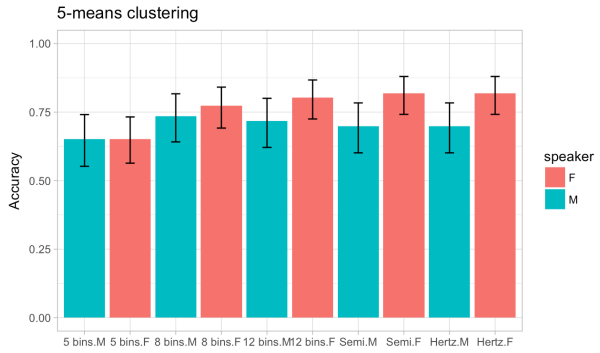

Figure 4: K-means clustering accuracy across measurement types

A pairwise test of proportions reveals no significant difference in accuracy between any conditions (Figure 4). With 5 bins, accuracy is lowest, which is not unexpected, since there may not be high enough resolution to discriminate between categories. There is also not a large difference between use of 8 and 12 bins for analysis, which suggests that 8 bins may be sufficient to capture most variation in this data set. Overall, we find that the use of a tone levels in clustering analysis is as able to distinguish between tonal categories as a continuous measure such as semitones or Hertz.

We also tested the ability to aggregate speaker data in clustering analysis. For this, we concatenated the data for both speakers and performed the same clustering algorithm with five clusters (Figure 5). We find that this method matches manual annotations with 76% accuracy. These results show that in this case, the ATLAS output is sufficiently standardized to allow for aggregation and analysis of data from speakers with very different pitch ranges. These clusters could help researchers hone in on possible tonemic categories when working with a previously unanalyzed language. Of course, we had prior knowledge of the correct number of clusters for the data (five), but experimenting with different numbers of clusters may help point the researcher in the right direction.

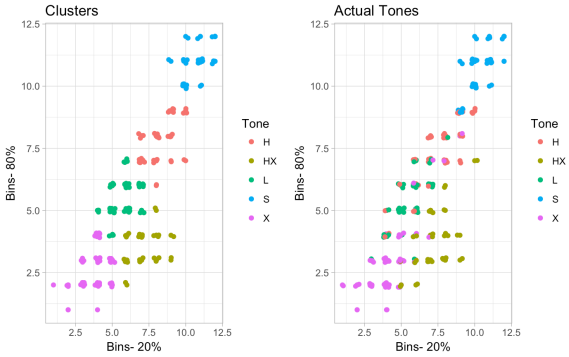

Figure 5: K-means clustering for aggregated speaker data (clusters vs. actual tonal categories).

The data used in this study were from a highly-controlled recording environment with a relatively small dataset. Further investigation of the use of ATLAS tone levels in naturalistic speech settings where f0 is more variable would help better illuminate the use of ATLAS's output. However, this case study shows that ATLAS discrete output can provide enough phonetic detail for analyses such as clustering and may allow for inter-speaker comparison.

5. Conclusion

To conclude, we have described a new toolkit for tonologists and those engaged in language documentation: ATLAS. ATLAS allows researchers to extract clean, quality pitch data from recordings, to normalize it, and to convert it to easily digestible tone levels representing a broad phonetic level of annotation. These annotations promote transparency in tonal annotation and can be created by anyone, even those unfamiliar or uncomfortable with tonal annotation. The annotations would even be of use in studying intonation in non-tone languages.

We would like to reiterate that ATLAS does not produce and does not replace the need for a phonological analysis of tone, but the discrete level outputs can be used as an aid in locating tone contrasts and in understanding the realization of phonemic tone categories. The raw f0 and normalized semitone outputs can also be used in further phonetic analysis.

In ongoing work, we are developing a fully automated version of ATLAS that removes the need for the TextGrid input, which currently is a bottleneck in scaling annotations up to full corpora of data. It is our hope that with a tool like ATLAS, we will no longer be faced with materials on tone languages that lack any indication of tone.

6. Acknowledgements

We would like to thank the Dartmouth College Neukom Institute and NSF-DEL grant BCS-1664335 for funding this research. Many thanks to Jim Stanford and audiences at ICLDC5 and APLL9 for helpful feedback on the tool.

7. References

[1] J.C. Rivierre. *Phonologie comparée des dialectes de l'extrême sud de la Nouvelle Calédonie.* SELAF, 1973.

[2] G.N. Clements, "The description of terraced-level tone languages," *Language* vol. 55, no. 3, pp. 536-448, 1979.

[3] J. Hualde and P. Prieto, "Towards and International Prosodic Alphabet IPrA," *Journal of Laboratory Phonology* vol. 7, no. 1, 2016.

[4] J. Wu, S. Zahorian, and H. Hu, "Tone recognition for continuous accented Chinese," in *International Conference on Acoustics, Speech, and Signal Processing,* pp. 7180-7183, 2013.

[5] W. J. Yang, J. C. Lee, Y. C. Chang, and H. C. Wang, "Hidden Markov model for Mandarin lexical tone recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36. no. 7, pp. 988–992, 1988.

[6] J. E. Cooper-Leavitt, "A computational classification of Thai lexical tones," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 2216–2216, 2016.

[7] T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng, and B. Mak, . "Tone recognition of isolated Cantonese syllables," *IEEE Transactions on Speech and Audio Processing,* vol. *3,* no. 3, pp. 204–209, 1995.

[8] O. Adams, T. Cohn, G. Neubig, & A. Michaud, "Phonemic transcription of low-resource tonal languages," *Proceedings of Australian Language Technology Association Workshop,* 53–50, 2017.

[9] R. Dockum, "Tone analysis in Southeast Asia: computational modeling and traditional methods," *Talk presented at the 26th Annual Meeting of the Southeast Asian Linguistics Society*, Manila, Philippines, 2016.

[10] S. Bird, "Automated tone transcription," 1994. Retrieved from http://arxiv.org/abs/cmp-lg/9410022

[11] D. Gibbon, U. Eno-Abasi, and M. Ekpenyong, "Problems and solutions in African tone language text-to-speech," *Proceedings of the Multiling 2006 Conference*, Stellenbosch, South Africa, 2006.

[12] R. J. Baken, *Clinical measurement of speech and voice.* Boston: College Hill Press, 1987.

[13] J. T. Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: an experimental approach to speech melody.* Cambridge: Cambridge University Press, 1990.

[14] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in M. Aronoff and R. Oehrle, eds, *Language Sound Structure*. Cambridge: MIT Press pp. 157-23, 1984.

[15] E. Ross, J. Edmondson, and G. Seibert, "The effect of affect on various acoustic measures of prosody in tone and non-tone languages: a comparison based on computer analysis of voice," *Journal of Phonetics,* vol. 14, no. 2, pp. 283-302, 1986.

[16] Y. Xu, "Understanding tone from the perspective of production and perception," *Language and Linguistics*, vol. 5, pp. 757-97, 2004.

[17] F. Nolan, "Intonational equivalence: an experimental evaluation of pitch scales," *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2003.