INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Automatic Analysis of Speech Prosody in Dutch

Na Hu[1], *Berit Janssen*[2], *Judith Hanssen*[3], *Carlos Gussenhoven*[4], *Aoju Chen*[1]

[1] Utrecht University, the Netherlands
[2] Digital Humanities Lab, Utrecht University, the Netherlands
[3]Avans University of Applied Sciences, the Netherlands
[4] Radboud University, the Netherlands

{n.hu, b.d.janssen, aoju.chen}@uu.nl, judithhanssen@gmail.com, c.gussenhoven@let.ru.nl

Abstract

In this paper we present a publicly available tool for automatic analysis of speech prosody (AASP) in Dutch. Incorporating the state-of-the-art analytical frameworks, AASP enables users to analyze prosody at two levels from different theoretical perspectives. Holistically, by means of the Functional Principal Component Analysis (FPCA) it generates mathematical functions that capture changes in the shape of a pitch contour. The tool outputs the weights of principal components in a table for users to process in further statistical analysis. Structurally, AASP analyzes prosody in terms of prosodic events within the auto-segmental metrical framework, hypothesizing prosodic labels in accordance with Transcription of Dutch Intonation (ToDI) with accuracy comparable to similar tools for other languages. Published as a Docker container, the tool can be set up on various operating systems in only two steps. Moreover, the tool is accessed through a graphic user interface, making it accessible to users with limited programming skills.

Index Terms: Dutch prosody, ToDI, FPCA, automatic prosody annotations

1. Introduction

Prosody (i.e. the melody of speech) is a critical aspect of spoken language. It provides the organizational structure of speech [1, 2] and is also vital to communication [3, 4, 5, 6]. Hence, implementing prosody in natural language systems such as speech synthesis and recognition is likely to augment system performance [7]. To this end, linguists and speech technologists working in the field have devoted much attention to the applications of Tones and Break Indices (ToBI) [8]. Taking a phonological perspective, ToBI considers prosody in terms of abstract prosodic events denoted by discrete prosodic labels, providing a symbolic representation of prosodic events. By spelling out a comprehensive set of rules, ToBI guides annotators throughout the annotation process. However, setting these labels manually is extremely labor-intensive (8-12 minutes per sentence per annotator) and costly in practice. Hence, an automatic solution is urgently needed.

Recent years have seen significant advances in the field of automatic annotation of English prosody following the ToBI notation. Systems have been developed using different machine learning techniques, including decision trees [9], neural networks [10, 11], and support vector machines [12, see [13] for an overview]. More specifically, AuToBI is the first publicly available tool for automatic annotation of main stream American English-ToBI (MAE-ToBI) labels, providing a ready-to-use solution for researchers/engineers in need of ToBI annotations [14]. The system performs six classification tasks: 1) pitch accent detection, 2) pitch accent classification, 3) intonational phrase detection, 4) intermediate phrase detection, 5&6) classification of phrase ending tones at both intonational and intermediate phrase boundaries. With logistic regression or support vector machine (SVM), the tool detects pitch accent with an accuracy of around 82.9%, and the boundaries of intonational phrase at 93.1% accuracy. The pitch accents were classified with a Combined Error Rate of 0.284.

However, using models of AuToBI to transcribe the prosody of another language yields mixed results. For example, using the AuToBI model to detect the pitch accent of Italian, French and German, [15] shows that the outcomes significantly differ across languages, pointing out the necessity of retraining the model using the data of the target language. Furthermore, by limiting prosody variations to a finite set of discrete labels, a ToBI-based tool cannot capture the rich variations in pitch properties [16]. As a complementary approach, Functional Principal Components Analysis (FPCA) describes the dynamics of pitch movements over the course of an utterance by representing the dominant modes of variations among input curves in terms of principal components (PCs) and calculates for each input curve the extent to which each PC is applied on it (the PC scores) [17]. Like conventional acoustic measures such as pitch and duration, the PC scores can be used in further statistical tests.

Taking into account the advantages of both phonological and functional approaches, the present project aims to develop the first publicly available tool that automatically analyses speech prosody – AASP in Dutch. Although it currently focuses on Dutch, it has the potential to be adjusted and used for other languages.

AASP performs two levels of analysis on prosody automatically: holistic and structural. Holistically, AASP performs a functional principal component analysis (FPCA) [17] on pitch curves, generating PC weights for individual curves for further statistical tests. Structurally, AASP predicts prosodic labels within the ToDI framework [18]. ToDI (Transcription of Dutch Intonation) is a transcription system of prosody designed specifically for standard Dutch. Sharing a similar philosophy with ToBI, ToDI analyzes Dutch prosody in terms of prosodic events such as pitch accents and prosodic boundaries, representing them by discrete labels. Different from MAE-ToBI [19], ToDI only defines one level of

http://dx.doi.org/10.21437/Interspeech.2020-2142

prosodic phrasing, instead of two. Moreover, the two systems differ in how pitch accents are analyzed. In MAE-ToBI, pitch accents captures both the contour leading towards the accented target and the contour leading off the accented target, whereas ToDI's pitch accents capture the contour leading off the accented target [20, 21, 22]. Table 1 shows the ToDI inventory, which consists of the location of the most salient part in a speech flow (pitch accent), pitch movements associated with the stressed syllable of a word (pitch accent types), the location of prosodic phrasal boundaries (prosodic boundaries), and pitch movements at the boundaries of an intonational phrase (prosodic boundary tones).

Table 1: *ToDI inventory.*

Prosodic events	**Decibels**
Pitch accent	Accented/unaccented
Pitch accent types	H*, !H*, H*L, !H*L, L*HL, L*, L*H, H*LH
prosodic boundary	Yes/no
prosodic boundary tones	%L, %HL, %H, H%, L%, %

AASP is freely distributed as a Docker container[1], which encapsulates the code and all its dependencies, so that the application runs quickly and uniformly in spite of differences between operation systems. The application and user manual of AASP can be downloaded from

`https://github.com/UUDigitalHumanitieslab/AASP`

To use the tool, users first need to install Docker on their own machine, download AASP to a local directory, and set up a Container for AASP in Docker, which can be done in only two steps by following the manual provided in the link. Different from existing tools of automatic prosody annotation, AASP is presented with a graphic user interface, which makes it friendly to users with limited programming skills.

2. System schematic

AASP consists of two independent analytical modules: AuToDI and FPCA. Figure 1 displays its schematic. First, users select the analysis of their choice. Then they are requested to specify a directory of files which should be analyzed. Only one type of analysis can be performed at a time, because the two modules require different input formats. Assuming that users need both analyses, they will need to execute the procedure twice.

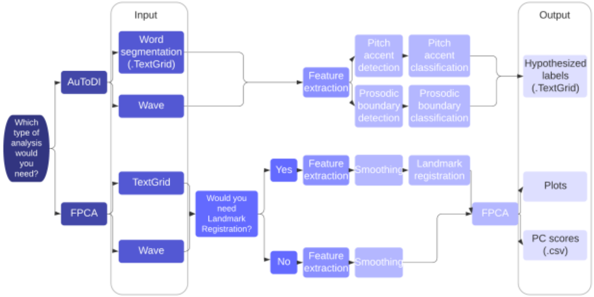
Which type of analysis would you need?
AuToDI
FPCA
Input
Word segmentation (.TextGrid)
Wave
TextGrid
Wave
Would you need Landmark Registration?
Yes
No
Feature extraction
Pitch accent detection
Pitch accent classification
Prosodic boundary detection
Prosodic boundary classification
Feature extraction
Smoothing
Landmark registration
Feature extraction
Smoothing
FPCA
Output
Hypothesized labels (.TextGrid)
Plots
PC scores (.csv)

Figure 1: *Schematic diagram of AASP*

2.1. AuToDI

AuToDI performs four classification tasks abiding to the ToDI conventions: 1) pitch accent detection; 2) pitch accent type classification; 3) prosodic boundary detection; 4) prosodic boundary tone classification.

AuToDI requires two types of user inputs: 1) audio files (in .wav format), which can be either a single long wave or a bunch of short waves contained in a folder; 2) the corresponding Praat TextGrid files containing word segmentations, which can be generated by external services provided by the OH-portal supported by CLARIN ERIC [23]. With the input data, AuToDI first extracts features including pitch, intensity and spectral information on a word level from the waveforms using the feature extraction module of AuToBI [14]. With the features, the tool first detects the locations of pitch accent and prosodic boundaries. Then, for words bearing pitch accents, the pitch accent types are predicted, and for words located at the prosodic boundaries, the boundary tones are predicted. Finally, the hypothesized labels are output to TextGrid files, which can be downloaded to a local directory of the users' choice.

2.2. FPCA

The operation of FPCA implements the workflow described in [17], which consists of three steps with the second step being optional: 1) smoothing, 2) landmark registration, 3) FPCA. First, the raw pitch curves are smoothed using B-spline interpolation and also rescaled according to the same time window. With this treatment, the micro-pitch variations irrelevant to experimental manipulation are smoothed out, resulting in a smooth curve. After smoothing, the curves can be further adjusted by aligning the time points of common internal landmarks shared by the curves, e.g. common phone boundaries [17] or common syllabic boundaries [24], if such common landmarks exist. By doing this, the pitch movements are aligned with respect to the common landmarks. Lastly, Functional PCA is conducted to extract the principal components as well as their weights for each curve. The aforementioned workflow is implemented as an R script, which is adapted from an open access script provided by [17].

FPCA requires two types of inputs from users: wave files and TextGrids containing boundaries of the domains of interest (DOI) and if applicable, locations of landmarks. The FPCA workflow starts with extracting relevant acoustic measures from the DOI, including f0 values at a step of 5 ms, the duration of the DOI, and if applicable, the duration of the region between landmarks, using Python implementations of Praat scripts. Then, the previously mentioned three steps (or two, if users do not need landmark registration) will be performed sequentially. In the process of smoothing, users will see a plot in a pop-up window showing the results of cross-validation of different combinations of the smoothing parameters: *k* and *lambda*. Users have to specify their choice of smoothing parameters so that the smoothing can proceed. The general principle is to pick from the combinations that yields similar smoothing results the combination that consists of the smallest *k* and the largest *lambda* (for details see [17]). After smoothing, the program will carry out landmark registration if users have opted for them and otherwise proceed with FPCA. Finally, the tool outputs PC weights of each PC for individual curves in a .csv file and plots showing how each PC manipulates the mean curve. These will be

[1] https://www.docker.com/resources/what-container

returned to users as a compressed folder, which can be saved at the location of their choice.

3. Methodology

The training set comprised of 4269 Dutch utterances collected in previous studies on Dutch intonation [25, 26]. The data set consists of 134 speakers of 6 Dutch dialects including Nijmeegs in South Guelderish, Zeelandic in Zuid-Beveland, Hollandic in Rotterdam and Amsterdam (AM) (all Low Franconian), West Frisian in Grou and Low Saxon in Winschoten. The utterances vary in sentence types (statements, yes/no questions and rhetorical questions) and also in the locations of pitch accent (IP-medial, IP-final). The utterances were scripted speech elicited from a dialogue setting, ensuring a fair representation of natural speech. The data were annotated by experienced annotators following ToDI conventions. Prosodic boundaries were annotated on the whole data set. Pitch accents and the associated accent types were annotated on a portion of the set (2600 utterances), resulting in a total number of 2868 instances distributed across the nine pitch accent types with an inter-annotator agreement of 0.998 (Cronbach's alpha) [25]. During annotation, annotators' uncertainty was denoted by "?", and an unrealized pitch target of a pitch accent was indicated by "()", resulting in additional 4 atypical classes including H*L?, H*(L), !H*(L) and !H*L?. Given that the main difference between an atypical instance and a typical one lies in phonetics rather than phonology and that the number of the atypical instances was small compared to the number of typical instances (e.g. H*L? only occurs 5 times in the whole set), the "?" cases were merged with the typical cases (e.g. H*L? converted to H*L), and the "()" cases were merged with the typical instances without the unrealized tone targets (e.g. H*(L) converted to H*). The number of instances for each accent type is shown in Table 2. The number of pitch accent types used in the analysis was reduced to seven by discarding H*LH because of scarce instances.

Table 2: *Number of instances per pitch accent class*

Pitch accent type	.Count
H*L	1014
H*	187
L*	44
L*H	617
H*LH	2
!H*L	178
!H*	34
L*HL	794

Acoustic features were extracted from the input files at the word level using the feature extraction module of AuToBI [14]. For each word, pitch, intensity, duration, and spectral information with mean, standard deviation, maximum, minimum, medium, slope, and range were obtained. Regarding pitch and intensity, the features were normalized using z-score to eliminate speaker differences. In addition, these features were also normalized relative to neighboring words. As in AuToBI, a unique set of features was constructed for each classifier (for more details of the feature extraction module see [14]).

The classifiers were trained using the Weka machine learning software [27] with 10-fold cross-validation. For each task, the performance of three types of classifiers were compared, the support vector machine (SVM) with linear kernel, logistic regression, and J48 (the java version of C4.5). We used Accuracy, Precision, and Recall as evaluation criteria to choose the model with the best performance.

4. Results

Table 3 shows the performance of SVM, logistic regression and J48 in the four classification tasks, pitch accent detection, pitch accent classification, prosodic boundary detection, and prosodic boundary tone classification. In general, SVM shows the highest accuracy in all four tasks, outperforming logistic regression and J48.

Table 3: *Accuracy of the three types of classifiers in the four detection and classification tasks*

Style Name	Classifier	Accuracy
Pitch accent detection	SVM	94.6%
	Logistic	86.9%
	J48	83.3%
Pitch accent classification	SVM	75.4%
	Logistic	68.2%
	J48	54.1%
Prosodic boundary detection	SVM	88.98 %
	Logistic	79.43%
	J48	72.6%
Prosodic boundary tone classification	SVM	84.7%
	Logistic	77.4%
	J48	78.6%

For pitch accent detection, SVM yielded the highest performance with an accuracy of 94.6% with a weighted average F-Measure 0.946 (Precision: 0.946, Recall = 0.946). Regarding pitch accent classification, SVM outperformed the other two classifiers with an accuracy of 75.4% with a weighted average F-Measure of 0.751 (Precision = 0.751, Recall = 0.754). In Table 4, the confusion matrix shows that H*L, L*HL, !H*L were relatively easy to classify with F-Measures ranging from 0.700 to 0.850, while L* exhibits the lowest F-measure 0.286. It is also clear from Table 4 that some accent types are easily confusable, suggesting some degree of similarity between pitch accent types. Specifically, L*H is misclassified as L*HL in 149 out of 617 instances, and L* is misclassified as L*H in 29 out of 44 instances. Also, !H*L and H*L were easily confusable – 44 out of 178 cases of !H*L were classified as H*L.

Table 4: *Confusion matrix of pitch accent types*

Classified as →	a	b	c	d	e	f	g	F-Measure
a = L*H	399	10	149	10	45	3	1	0.627
b = L*	29	10	0	0	4	0	1	0.286
c = L*HL	121	0	647	3	20	0	3	0.810
d = H*	32	1	0	101	43	7	3	0.591
e = H*L	66	5	7	30	872	1	33	0.850
f = !H*	2	0	0	7	9	14	2	0.437
g = H*L	6	0	0	4	44	5	119	0.700

With respect to prosodic boundaries, the results show that the locations of prosodic boundaries can be detected by SVM with an accuracy of 88.98% with a weighted average F-Measure of 0.882 (Precision: 0.885, Recall = 0.890), and the tone types at prosodic boundaries can be classified with an accuracy of 84.7%.

With regard to FPCA, the tool outputs PC weights for raw curves in a .csv file, which can be read into statistical software, such as R or SPSS, for further statistical tests. For examples of prosody research using FPCA output, the readers are referred to [17, 28, 29, 30].

5. Discussion

In general, the results of Automatic ToDI label predictions are consistent with the results reported in previous studies. That is, pitch accent detection and prosodic boundary detection are relatively easy, performing with high accuracy, whereas pitch accent classification is the hardest, showing the lowest accuracy of all. Specifically, our tool predicts the location of pitch accent with an accuracy of 94.8% using SVM, comparable to the 90% reported in [11]. And for prosodic boundary detection, our tool shows an accuracy of 88.98% using SVM. One reason for the yielded high accuracy is that the training set contained reliable annotations (high inter-annotator agreement). With regard to pitch accent classification, the classifier with the best performance, which is SVM, shows an accuracy of 75.4%, slightly better than the 70.8% reported in [13], which adopted the same number of pitch accent types as the current tool. Note that our result is only based on seven types of pitch accent, excluding H*LH, which only had two instances in the original training set and was therefore discarded. In fact, although its pattern is phonologically distinguishable from other types, H*LH's are restricted to pre-final accent locations and may be even more infrequent in natural speech. The difficulty in predicting pitch accent types may be partially due to the scarce data of certain accent types. In our training set, the number of instances across types was unbalanced (e.g. 34 !H* vs. 1014 H*L). Further, the difficulty in classifying pitch accents might lie in the fact that the boundaries between accent types are naturally fuzzy. To deal with the resemblances between certain pitch accent types, some researchers divided the original accent types into groups based on the extent to which one type of pitch accent was similar to the other, and consequently, gained an improvement in model performance. For example, by grouping ToBI accents into high (H*, L+H*, H+!H*), downstepped (!H*, L+!H*), and low (L* and L*+H), [31] gained an accuracy of 81.3%, and [32] achieved an accuracy of 87.17% using ensemble learning methods, generally higher than studies adopting original categories. Besides more training data and a grouping criterion, the performance of pitch accent classification might be improved by adopting new features, especially the features that portray the characteristics of pitch contours. In this regard, FPCA shows great potential as it captures the main mode of variations among the contour shapes. This will be further investigated in our future work.

6. Conclusions

We presented an initial version of AASP, which is the first publicly available tool to automatically annotate Dutch prosody. It enables users to perform two types of analysis: ToDI and FPCA. With respect to ToDI, the tool performs four tasks including pitch accent detection, pitch accent classification, prosodic boundary detection, and prosodic boundary tone classification. Using SVM, the tool performs with accuracy comparable to similar tools of other languages. Regarding FPCA, AASP outputs the weights of principal components in a .csv file, which can be directly used for further statistical tests.

The tool is packaged as a Docker container that can run on a wide range of operating systems. Also, it comes with a graphical user interface to ensure ease of use for users with limited programming skills. Future work will explore new features such as the PC weights generated by Functional Principal Component Analysis (FPCA) in order to examine if they can improve the accuracy of pitch accent classification.

7. Acknowledgements

We would like to thank Andrew Rosenberg for his help and Rick Janssen for his support throughout this project, José de Kruif, Jelte van Boheemen and Sheean Spoel for their input in the early stage of this project, and Arjan van Hessen for his help with software for automatic transcription of spoken Dutch. This project is supported by a Research IT innovation fund of Utrecht University to Aoju Chen.

8. References

[1] M. Nespor and I. Vogel, *Prosodic Phonology*. Dordrecht, the Netherlands: Foris., 1986.

[2] M. E. Beckman, "The parsing of prosody," *Lang. Cogn. Process.*, vol. 11, no. 1–2, pp. 17–68, 1996.

[3] J. Cole, "Prosody in context: a review," *Lang. Cogn. Neurosci.*, vol. 30, no. 1–2, pp. 1–31, 2015.

[4] A. Cutler, D. Dahan, and W. Van Donselaar, "Prosody in the Comprehension of Spoken Language: A Literature Review," *Lang. Speech*, vol. 40, no. 2, pp. 141–201, 1997.

[5] P. Prieto, "Intonational meaning," *Wiley Interdiscip. Rev. Cogn. Sci.*, 2015.

[6] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in communication*, P. R. Cohen, J. L. Morgen, and M. E. Pollack, Eds. MIT Press, Cambridge, MA; London. Lewis., 1990.

[7] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," *Proceedings of ISCA Workshop Prosody in Speech Recognition and Understanding*, 2001, pp. 13–16.

[8] M. E. Beckman, "ToBI□: A standard for labeling English prosody," *Proceedings of the 1992 International Conference on Spoken Language Processing*, 1992, pp. 12–16.

[9] C. W. Wightman and M. Ostendorf, "Automatic Labeling of Prosodic Patterns," *Proceedings of IEEE Transactions on Speech and Audio Processing*, 1994.

[10] C. González-ferreras, D. Escudero-mancebo, C. Vivaracho-pascual, and V. Cardeñoso-payo, "Improving automatic classification of prosodic events by pairwise coupling," *Proceedings of IEEE Transactions on Audio, Speech, and Language Processing*, 2012, vol. 20, no. 7, pp. 2045–2058.

[11] C. Ni, W. Liu, and B. Xu, "Automatic Prosodic Events Detection by Using Syllable-Based Acoustic , Lexical and Syntactic Features," in *INTERSPEECH*, 2011.

[12] J. H. Jeon and Y. Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4565–4568.

[13] D. Escudero-Mancebo, C. González-Ferreras, C. Vivaracho-Pascual, and V. Cardeñoso-Payo, "A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling," *Comput. Speech Lang.*, 2014.

[14] A. Rosenberg, “AuToBI - A tool for automatic ToBI annotation,” *Proceedings of INTERSPEECH*, 2010, pp. 146–149.

[15] A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg, “Cross-language prominence detection,” *Proceedings of Speech Prosody,* 2012, vol. 1, pp. 278–281.

[16] G. Lohfink, A. Katsika, and A. Arvaniti, “Variability and category overlap in the realization of intonation,” *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019.

[17] M. Gubian, F. Torreira, and L. Boves, “Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts,” *J. Phon.*, vol. 49, pp. 16–40, 2015.

[18] C. Gussenhoven, “Transcription of Dutch Intonation,” in *Prosodic Typology: The Phonology of Intonation and Phrasing*, 2010.

[19] C. Gussenhoven, “Correction: Analysis of Intonation: The Case of MAE_ToBI,” *Lab. Phonol.*, 2016.

[20] C. Gussenhoven, “Semantic judgments as evidence for the intonational structure of Dutch,” *Proceedings of Speech Prosody*, 2008.

[21] J. Hanssen, J. Peters, and C. Gussenhoven, “Prosodic effects of focus in Dutch declaratives,” *Proceedings of Speech Prosody*, 2008.

[22] A. Chen, “What’s in a rise: Evidence for an off-ramp analysis of Dutch intonation,” *Proceedings of ICPhS*, 2011.

[23] C. Draxler, V. den H. H., A. Van Hessen, S. Calamai, L. Corti, and S. Scagliola, “A CLARIN Transcription Portal for Interview Data.,” *Proceedings of LREC*, 2020.

[24] G. Turco and M. Gubian, “L1 prosodic transfer and priming effects: A quantitative study on semispontaneous dialogues,” *Proceedings of Speech Prosody*, 2012.

[25] J. Hanssen, “Regional variation in the realization of intonation contours in the Netherlands,” Ph.D. dissertation, Radboud University, 2017.

[26] N. Hu, A. Chen, H. Quené, and T. Sanders. (in preparation). “The role of prosody in expressing subjective and objective causality in Dutch discourse.”

[27] E. Frank, M. A. Hall, and I. H. Witten, “The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques,” *Morgan Kaufmann, Fourth Ed.*, 2016.

[28] A. Chen and L. Boves, “What’s in a word: Sounding sarcastic in British English,” *J. Int. Phon. Assoc.*, vol. 48, no. 1, pp. 57–76, 2018.

[29] M. Gubian, L. Boves, and F. Cangemi, “Joint analysis of f0 and speech rate with functional data analysis,” *Proceedings of ICASSP*, 2011.

[30] O. Jokisch, T. Langenberg, and G. Pintér, “Intonation-based classification of language proficiency using FDA,” *Proceedings of Speech Prosody*, 2014.

[31] G. A. Levow, “Context in multi-lingual tone and pitch accent recognition,” in *INTERSPEECH*, 2005.

[32] X. Sun, “Pitch accent prediction using ensemble machine learning,” *Processings of ICSLP*, 2002, pp. 16–20.