WHAT'S THAT *PHTHONG*? AUTOMATED CLASSIFICATION OF DIALECTAL MONO- AND STANDARD DIPHTHONGS

Simon Oppermann & Beat Siebenhaar

Universität Leipzig
simon.oppermann@uni-leipzig.de, siebenhaar@uni-leipzig.de

ABSTRACT

In East Central German, there’s an opposition in the reflexes of Middle High German *ei*, *ou*, *öu*: In Standard German, these diphthongs are shifted to [aɪ̯, aʊ̯, ɔɪ̯] whereas in the dialects they’ve been monophthongised to [eː, oː, eː] respectively. However, dialectal monophthongs are increasingly replaced by standard diphthongs. This paper uses publicly available data from non-professional speakers recorded over the past 20 years to examine (supposed) lifespan variation in *phthong* use. Combatting this plethora of data by auditory categorisation alone is not feasible. Thus, a procedure to automate the distinction between mono- and diphthongs for larger datasets is proposed: Relevant segments are force-aligned and formant tracks are calculated automatically. Their DCT coefficients and temporal parameters are then used to train multiple random forests. With a classification accuracy of over 95% this scalable model promises sufficiently accurate results for the analysis of lifespan change in *phthong* realisations.

Keywords: DCT, random forest, diphthong, East Central German, lifespan change.

1. INTRODUCTION

In our research project, we are interested in examining linguistic (in-)stability across the lifespan in the use of multiple East Central German (ECG) variables. In this paper, the focus will only be on the realisation of Middle High German (MHG) *ei* and *ou*. As the available dataset currently consists of more than 420 hours of footage, this paper proposes a procedure for automating the distinction between monophthongs and diphthongs using random forest models.

1.1. Automatic distinction of mono- and diphthongs

In sociophonetics, the standard practice to analyse vowels acoustically is the calculation of the first and second formants. These static measurements of F1 and F2 at the temporal midpoints of each vowel, commonly used for distinguishing monophthongs, lack the ability to account for the diphthongs’ dynamic nature, which is manifest as a temporal systematic change in formant values. In recent years, several different methods have been proposed to quantify the formant paths of diphthongs. Instead of measuring just the vowel centre, the measurement is extended to multiple time points. Here, time points can either be chosen relative to the vowel length, e.g. at discrete points spaced equally around the midpoint of each vowel [1, 2, 3], or in absolute steps, e.g. every 2.5 milliseconds [4]. Various methods have been proposed to utilise these discrete values to represent the dynamic formant trajectories: polynomial functions, additive models, target-locus scaling, vector-based measurements, and discrete cosine transformations [2]. Polynomial functions [5] and additive models [4, 6, 7] fit “curves to the sampled formant frequencies over time” [3]. Although “they have the advantage of not forcing a parameterisation on trajectory shapes” [3], they are difficult to “generalize across different datasets” [3]. Vector-based approaches [2, 8] measure the Euclidian distance between two data points and/or calculate the angle of these vectors. Despite relying on little spectral information, they have proven to be quite effective [3]. Nonetheless, classification results can be even further improved by representing trajectory shapes with discrete cosine transformations [2].

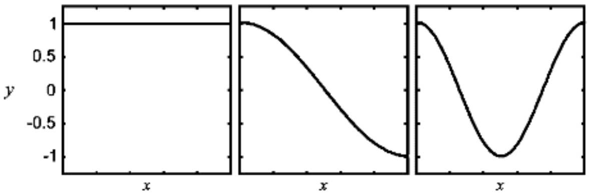

Figure 1: Zeroth, first and second DCT basis functions. Taken from [8], edited by the first author.

Discrete cosine transformations (DCT) [1, 3, 10, 11, 12] work by expressing an input signal sequence as a sum of weighted cosine functions with different frequencies. “The curve parameterization is based on a trajectory’s mean plus ½ cosine multiples, each with amplitudes representing deviations from this mean” [3]. The amplitudes of these cosine functions are referred to as DCT coefficients. Figure 1 shows that the zeroth DCT coefficient DCT-0 correlates with the mean of the input signal, DCT-1 (half of a cosine), with the slope's magnitude and direction and DCT-2 (a full cosine) with the overall curvature [3]. Several studies use these first three DCT coefficients to successfully describe vowel-inherent spectral change [3, 10, 11, 12], while DCT-0 and DCT-1 are the most important [1, 3, 9].

1.2. Lifespan variation and change

In the past two decades, an ever-growing number of studies have been published that more closely examine lifespan stability, variation, and change in several languages (see e.g. the edited volumes [13, 14]). For German, there are only a few studies to date [4, 15, 16, 17, 18]. All these studies show a great amount of inter-individual variation: While the greater part of individuals show linguistic stability, a few others show considerable change over time. E.g. in [4], the majority of speakers change their language over the course of their lifespan, following the community-wide trend toward variants closer to the standard. Only two individuals reveal a retrograde change which is explained by retirement and a pronounced sense of dialect identity. The result that most speakers change their language over their lifespan contradicts those findings, that stability is the predominant individual pattern [19]. However, in a situation with a strong dialect levelling as in Swabia [4], it is likely that most community members conform to the trend.

The majority of these studies only work with two different recording dates [4, 15, 16, 17, 18], and the differences they find at these two stages are interpreted as lifespan change. Studies that look at more time stamps [21, 22, 23] reveal more complex structures, as the realisations of different variants often seem to oscillate, which can be interpreted as a somewhat stable variation. If only two points in time are considered within such oscillating patterns, these differences are overinterpreted or misinterpreted as change. Against this methodological background, a denser temporal grid for the investigation of lifespan change (or variation) would be preferable [21].

1.3. The linguistic situation in East Central German

The data for this paper was recorded in the ECG area, where on the one hand the traditional dialects have already been largely lost and their function is now filled by a regional vernacular – in contrast to many other German areas. On the other hand, the regional realisation of the standard contains many of these vernacular features so that standard and vernacular tend to be quite close to each other [24, 25]. One of the features considered typical of ECG dialects is the realisation of the MHG diphthongs *ei*, *ou*, and *öu* as monophthongs [eː], [oː] and [eː] respectively, thus preserving a phonological contrast that has been lost in New High German (NHG). At present, an abandoning of the monophthongal realisations in the regiolect [24] and thus a trend towards the standard diphthongs can be observed.

MHG	*î*	*ei*	*û*	*ou*	*iu*	*öu*
ECG	aɪ̯	eː	aʊ̯	oː	ɔɪ̯	eː
NHG	aɪ̯	aɪ̯	aʊ̯	aʊ̯	ɔɪ̯	ɔɪ̯

Table 1: Merger of six MHG mono- and diphthongs to three NHG diphthongs. In ECG dialects, the phonological contrast is preserved.

2. DATA

The data basis for this study is the German zoo docusoap "Elefant, Tiger & Co." which follows the everyday life of the zoo animals and the work of Leipzig Zoo employees and has been broadcast weekly since 2003. The employees are accompanied during their work, they speak directly to the TV audience, to the animals, and to each other, while acting entirely without a script. Most speakers' educational, occupational, and socio-professional statuses remain consistent throughout the series. As of the submission of this paper, over 1000 episodes have been broadcast, each approximately 25 minutes long. For the current study, 133 of these episodes are divided into four evenly spaced periods: 2003/2004, 2008/2009, 2013/2014, and 2018/2019. In total, more than 20 regular interviewees have made appearances in at least three of these periods. Three of the most prominent speakers are selected for this paper.

3. METHOD

Each episode is transcribed orthographically in Praat [27]. These transcripts are then automatically force-aligned using WebMAUS [28] and other BAS web services [29, 30]. For the three selected speakers, 258,515 segments were obtained, 8,786 of these being either /aɪ̯/, /aʊ̯/, or /ɔɪ̯/. F1 and F2 values are calculated at 31 equally spaced time points (20%, 22% … 80% of total vowel duration) in Praat and z-normalised in R (v4.2.2) [31] using the package vowels [32]. Additionally, the duration of every segment is extracted. The NHG diphthong in each lexeme is then matched to the corresponding MHG vowel. 4,683 tokens (= 53.3%) correspond to MHG *î*, *û*, *iu*, which are realised as the same diphthongs in standard and dialect (see Table 1). Those that go back to MHG *ei*, *ou*, *öu* add up to 3,662 tokens (= 41.68%). 262 tokens (= 2.98%) correspond to neither of those historical categories. MHG *öu* is by far the rarest with only 83 tokens overall.

In the present, data only diphthongal realisations [ɔɪ̯] can be found. Hence all tokens for MHG *öu* will be excluded from further analysis.

Due to the large size of the entire corpus, neither the force alignment nor the formant paths can be corrected manually. To tidy the data automatically and mitigate the influence of potential noise in the process, tokens with a median F1 less than 200 Hz or larger than 1000 Hz, with a median F2 less than 800 Hz or larger than 2500 Hz, or those longer than 400 ms are excluded [4]. In Table 2 all remaining tokens for the actual analysis are listed, separated by speaker and period.

	MHG	Period			
	vowel	03/04	08/09	13/14	18/19
Speaker 1	*ei*	323	380	253	316
	ou	242	256	162	195
Speaker 2	*ei*	58	29	24	227
	ou	49	12	13	94
Speaker 3	*ei*	307	70	55	173
	ou	138	28	39	84

Table 2: MHG vowel tokens per speaker per period.

For each token, the first three DCT coefficients for F1 and F2 are calculated in R using the package emuR [33]. Together with the log10-transformed duration values these DCT coefficients are used as predictors for a random forest model, calculated with the R-package randomForest [34]. Random forest models have proven to be more accurate than regression models in separating multivariate linguistic data [35], despite being somewhat hard-to-interpret black boxes. They work by fitting multiple decision trees to data, but for each tree not only randomising the specific samples using bootstrap aggregation (bagging) but also changing the subsets of the predictor variables [35]. The final prediction is the average of all individual trees' predictions.

The specific random forest used to categorise automatically all tokens into either mono- or diphthongs is trained on a subset of tokens ($n = 324$), that previously have been auditorily categorised by the two authors and two student assistants. Tokens assigned being neither mono- nor diphthongs, mostly due to wrong alignment or phonetic reduction, have been excluded from this subset.

4. RESULTS

To achieve the most accurate classification, the random forests were calculated in three different ways: A) Should the model distinguish between MHG vowels? After all, different phonemes correspond to different formant tracks. B) Should the formant data be normalised? Z-normalisation is common when comparing formant data of different speakers. However, previous test runs with smaller samples achieved better results with non-normalised data. C) Should the model distinguish between different speakers? Here, only three speakers are analysed, but ideally the model should work for all the other additional speakers to be analysed later.

± MHG	± Norm.	± Speaker	Accuracy
+	+	+	.909
+	+	−	.896
+	−	+	**.976**
+	−	−	.969
−	+	+	.877
−	+	−	.883
−	−	+	.944
−	−	−	.938

Table 3: Highest classification accuracy scores of different random forest models, each calculated with unique parameters explained above.

Each of the eight configurations in Table 3 was tested on 40 different random forest model settings, each with different numbers of decision trees (ntree = 50, 100 … 2000) and repeated 100 times. The data was split 50/50 into training and test datasets. As there were eight or nine independent variables, depending on whether speaker information was included, the number of variables tried at each split (mtry) was limited to three, as recommended [35]. Where MHG vowels were differentiated, the best classification accuracy scores for MHG *ei* and MHG *ou* respectively were averaged. Table 3 shows that the highest classification scores were achieved with random forest models that were separated by MHG vowel, used non-normalised data, and differentiated between different speakers. The best-performing models achieved an accuracy score of .964 for MHG *ei* and .988 for MHG *ou*, resulting in an average accuracy of .976. As those models were trained at a very specific subset of the original data, they were then used to classify the *phthongs* with different data subsets. After 100 iterations each, the model for MHG *ei* achieved accuracies between .964 and 1 with a median of .976, and the model for MHG *ou* between .988 and 1 with a median of .988.

According to the variable importance scores (see Table 4) the first two DCT coefficients for F1 and F2 generally were the most important variables in both random forest models, contributing best to the homogeneity at the nodes and leaves (= Gini) and leading to the biggest drop in accuracy if they were excluded (= Acc.).

Variable	Acc. *ei*	Gini *ei*	Acc. *ou*	Gini *ou*
k0_f1	8.351	6.724	46.042	8.205
k0_f2	7.859	6.896	4.649	2.424
k1_f1	5.347	6.207	5.283	3.359
k1_f2	5.278	5.698	1.427	2.116
k2_f1	3.171	3.684	2.461	2.326
k2_f2	-0.197	2.710	-6.218	1.646
length	7.013	6.537	-0.091	1.993
speaker	1.184	0.828	4.873	0.718

Table 4: Variable importance scores (mean decrease in accuracy and mean decrease of Gini coefficient) of the best performing random forest models for MHG *ei* and *ou* respectively.

Length is the third most important variable for classifying the realisations of MHG *ei*, but one of the least important ones for MHG *ou*. Information about the speaker also turned out not to be too essential for classification, an observation supported by the fact that those random forests excluding speaker information altogether achieved only minimally worse accuracy scores than those that did not (see Table 3). In a final step, the two best-performing random forests – one for MHG *ei* and one for MHG *ou* – were used to analyse the usage of dialectal monophthongs across the three individual speakers' lifespans.

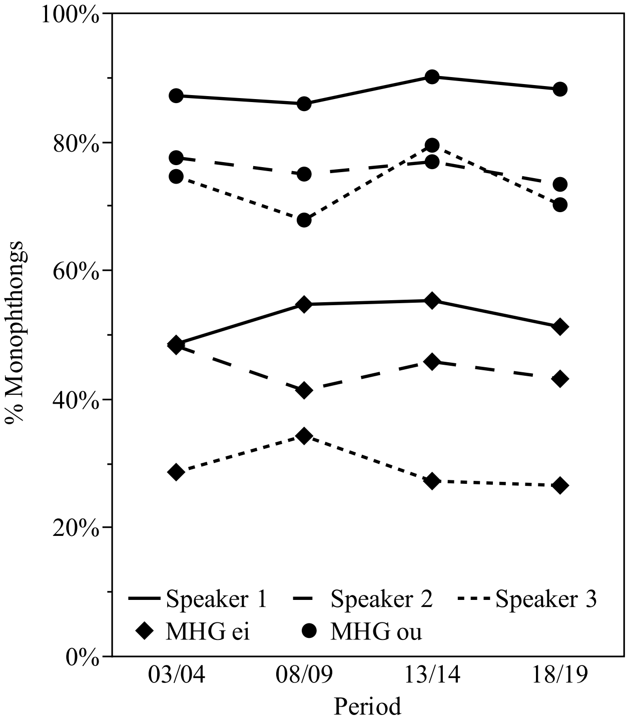

Figure 2: Percentage of monophthongs per speaker per period. Phonemes were classified automatically using the best-performing random forest models for MHG *ei* and *ou* respectively.

Overall, all speakers seem to show stability in their use of *phthongs* but tend to use a higher percentage of monophthongal realisations of MHG *ou* than of MHG *ei* (see Figure 2). Where variation does occur, it is of an oscillatory nature, increasing and decreasing alternately. The envelope of variation across the lifespan is quite low, remaining comfortably in single digits, except for MHG *ou* of Speaker 3. These small variations are reflected in the results of subsequent Pearson's chi-square tests at $\alpha = .05$, which show no statistically significant differences for each speaker and each phoneme ($p > .25$ in all cases). In contrast to this low level of intra-speaker variation, the overall inter-speaker differences are significantly distinct for MHG *ei*, χ^2 (2, N = 2,215) = 94.36, $p < .001$ and MHG *ou*, χ^2 (2, N = 1,312) = 39.21, $p < .001$.

The slight oscillations of the lifespan trajectories, not exhibiting statistically significant differences, could very well be influenced by comparatively low token numbers. Other factors possibly contributing to the varying use of mono- or diphthongs include situation-specific performance effects, as the speakers are recorded in different communicative situations, as well as the frequency of the respective words since in ECG the monophthongisation tends to be highly lexicalised.

5. DISCUSSION

Methodologically, random forest models with the three first DCT coefficients of F1 and F2 and duration of force-aligned segmentations can reach high accuracies in distinguishing monophthongs and diphthongs – at least in the ECG area. Hence this procedure has proven to be valuable for analysing the use of these *phthongs* in larger datasets.

Using this method, the analysis of lifespan variation of three adult ECG speakers reveals relative stability, sometimes slightly oscillating around their individual means. However, the speakers are different in their individual realisation of mono- and di-phthongs, resulting in a considerable inter-speaker variation. Their quasi-stable use of mono-phthongisation mirrors their stable educational, occupational, and socio-professional circumstances.

The general change within the community as reported in [24] is not reflected in an individual change, supporting findings of general stability as one of several possible patterns of individual linguistic variation across the lifespan [19]. Nonetheless, the corpus allows for further individuals to be analysed over their lifetime and for apparent time analyses at different time points with a dataset of more than 100 speakers. Thus, a deeper insight into the correlation between language change and individual variation in a longitudinal study can be expected.

6. REFERENCES

[1] Elvin, J., Williams, D., Escudero, P. 2016. Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *J. Acoust. Soc. Am* 140(1), 576–581.

[2] Tanner, J., Sonderegger, M., Stuart-Smith, J. 2022. Multidimensional acoustic variation in vowels across English dialects. In: Nicolai, G., Chodroff, E. (eds) *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. ACL, 72–82.

[3] Williams, D., Elvin, J., Escudero, P., Gafos, A. I. 2019. Multidimensional variation in English diphthongs. *Proc. 19th ICPhS* Melbourne, 587–591.

[4] Beaman, K. V. 2021. *Coherence in Real- and Apparent-Time: A sociolinguistic variationist investigation of language change in Swabia*. PhD Thesis. Queen Mary University of London.

[5] Themistocleous, C. 2017. Dialect classification using vowel acoustic parameters. *Speech Communication* 92. 13–22.

[6] Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., Groake, E. 2019. Dialect variation in formant dynamics: the acoustics of lateral and vowel sequences in Manchester and Liverpool English. *J. Acoust. Soc. Am* 145(2), 784–794.

[7] Renwick, M. E. L., Stanley, J. A,. 2020. Modeling dynamic trajectories of front vowels in the American South. *J. Acoust. Soc. Am* 147(1), 579–595 .

[8] Olsen, R. M., Olsen, M., and Renwick, M. E. L. 2018. The impact of sub-region on /ai/ weakening in the U.S. South. *Proc. Mtgs. Acoust.* 31(1), 060005

[9] Morrison, G. S. 2013. Theories of vowel inherent spectral change. In: Morrison, G. S., Assmann P. F. (eds), *Vowel Inherent Spectral Change*. Springer, 31–48.

[10] Watson, C. I., Harrington, J. 1999. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *J. Acoust. Soc. Am* 106(1), 458–468.

[11] Williams, D., Escudero, P. 2014. A crossdialectal acoustic comparison of vowels in Northern and Southern British English. *J. Acoust. Soc. Am* 136(5) 2751–2761.

[12] Wolfswinkler, K., Harrington, J. 2021. The influence of Standard German on the vowels and diphthongs of West Central Bavarian. *J. Int. Phon. Assoc.* 1–33

[13] Beaman, K. V., Buchstaller, I. (eds) 2021. *Language Variation and Language Change Across the Lifespan. Theoretical and Empirical Perspectives from Panel Studies*. Routledge.

[14] Wagner, S. E., Buchstaller, I. (eds) 2017. *Panel Studies in Language Variation and Change*. Routledge.

[15] Siebenhaar, B. 2002. Sprachwandel von Sprachgemeinschaften und Individuen. In: Häcki Buhofer, A. (ed), *Spracherwerb und Lebensalter*. Francke, 313–325.

[16] Bausch, K.-H. 2000. Dialektologie und interpretative Soziolinguistik am Beispiel des Sprachwandels im Rhein-Neckar-Raum. In: Stellmacher, D. (ed), *Dialektologie zwischen Tradition und Neuansätzen*. Steiner, 78–98.

[17] Bülow, L., Vergeiner, Ph. C. 2021. Intra-individual variation across the lifespan: Results from an Austrian panel study. In: *Linguist. Vanguard* 7(s2), 1–11.

[18] Ruge, J. 2011. Veränderungen im Dialektgebrauch derselben Sprecher innerhalb von drei Jahrzehnten. In: Glaser, E., Schmidt, J. E., Frey, N. (eds), *Dynamik des Dialekts — Wandel und Variation.* Steiner, 287–300.

[19] Sankoff, G. 2019. Language Change Across the Lifespan: Three Trajectory Types. *Language* 95(2), 1–36.

[21] Bowie, D. 2019. Individual variation in the development of the Western Vowel System of Utah. *Linguist. Vanguard* 5(2), 20180020.

[22] Reubold, U., Harrington, J., Kleber, F. 2010. Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Commun.* 52, 638–651.

[23] Reubold, U., Harrington, J. 2017. The Influence of Age on Estimating Sound Change Acoustically From Longitudinal Data. In: Wagner, S. E., Buchstaller, I. (eds), *Panel Studies in Language Variation and Change*. Routledge, 129–151.

[24] Rocholl, M. J. 2015. *Ostmitteldeutsch – eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum*. Olms.

[25] Siebenhaar, B. 2019. Ostmitteldeutsch: Thüringisch und Obersächsisch. In: Herrgen, J., Schmidt, J. E. (eds): *Deutsch: Sprache und Raum – Ein Internationales Handbuch der Sprachvariation*. De Gruyter, 407–435.

[27] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer. Computer program. Version 6.3.02. http://www.praat.org/ (accessed 16.11.2022)

[28] Schiel, F. 2015. A Statistical Model for Predicting Pronunciation. *Proc. 18th ICPhS* Glasgow. Paper number 195.

[29] Poerner, N., Schiel, F. 2018: A Web Service for Pre-segmenting Very Long Transcribed Speech Recordings. *Proc. LREC* Miyazaki, 2860–2865.

[30] Reichel, U. D. 2012. PermA and Balloon: Tools for string alignment and text processing. *Proc. Interspeech*. Portland, OR. Paper number 346.

[31] R Core Team. 2022. R: A language and environment for statistical computing. https://www.R-project.org/ (accessed 16.12.2022).

[32] Kendall T., Thomas, E. R. 2018. vowels: Vowel Manipulation, Normalization, and Plotting. R package version 1.2-2. https://CRAN.R-project.org/package=vowels (accessed 14.12.2022).

[33] Winkelmann, R., Harrington, J., Jänsch, K. 2017. EMU-SDMS: Advanced speech database management and analysis in R. In: *Comput. Speech Lang.* 45, 392–410.

[34] Liaw A., Wiener, M. 2002. Classification and Regression by randomForest. *R News* 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/ (accessed 14.12.2022).

[35] Tomaschek, F., Hendrix, P., Baayen, R. H. 2018. Strategies for addressing collinearity in multivariate linguistic data. *J. Phon.* 71, 249–267