

Characterising Formant Tracks in Viennese Diphthongs for Forensic Speaker Comparison

Ewald Enzinger¹

¹*Acoustics Research Institute, Austrian Academy of Sciences, Vienna, 1040, Austria*

Correspondence should be addressed to Ewald Enzinger (ewald.enzinger@oeaw.ac.at)

ABSTRACT

This study evaluates methods that capture time-dynamic properties of diphthongs produced by speakers of Viennese German for application in a forensic setting. Polynomials, discrete cosine transform and B-splines along with experimental features based on bent-cable regression models were used to characterise the first three formant tracks of two /æ/ diphthong segments. The resulting coefficients were in turn used as parameters in a speaker discrimination procedure based on likelihood ratios which were calculated using a multi-variate kernel density formula (MVKD). A comparison of the achieved performance based on cross-validation is presented in terms of equal error rate (EER) and the log-likelihood ratio cost metric as well as DET plots.

1. INTRODUCTION

In the domain of forensic speaker comparison, acoustic features used to capture speaker-specific properties are most commonly static in nature. Methods based on statistics of measurements from recordings, e.g. long term formant distributions [5], as well as automatic speaker verification systems [16, 2] model speakers based on these parameters. However, information about the temporal structure, the progression of these parameters over time, and their phonetic context is largely neglected.

Approaches focusing on these characteristics have recently been tested and evaluated for their capacity to discriminate between different speakers. Their common objective is to provide parametric representations of time-dynamic properties of speech segments. Quadratic and cubic polynomials [8, 9, 12] fitted to formant contours as well as components derived by applying a discrete cosine transform (DCT) to the same trajectories [14, 13] have been suggested for the use in forensic speaker comparison.

This study evaluates these two methods proposed in the literature in addition to two experimental models to characterise the progression of the formant trajectories within diphthongal segments.

2. PARAMETRIC FORMANT TRAJECTORY REPRESENTATIONS

The evaluation presented in this paper investigates the use of four different parametric representations of formant trajectories of diphthongs for the purpose of discriminating speakers. The first method is based on quadratic and cubic polynomials. Equation 1 shows the generic form of a polynomial function.

$$y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_k x^k \quad (1)$$

The coefficient values α_k determine the shape of the polynomial and are calculated by least-squares fitting. Figure 1 shows an example.

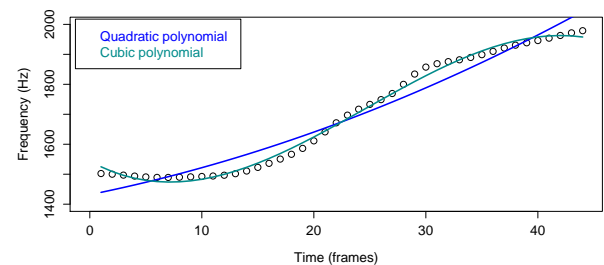


Fig. 1: Polynomial curves fitted to F2 of /æ/ in *Kreide*

The second representation is obtained by applying a *discrete cosine transform (DCT)* to the formant values. Equation 2 shows the formal definition of the transform (*DCT-II*).

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right] \quad k = 0, \dots, N-1. \quad (2)$$

The first three to four components are used as parameters for speaker comparison. Figure 2 shows the curves derived from the inverse DCT based on these values.

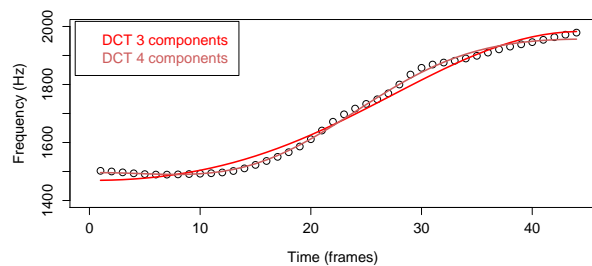


Fig. 2: DCT based on F2 of /æ/ in Kreide

Two further new representations are explored in this study. *B-splines*, i.e. pairwise polynomials, are fitted to the formant trajectories in the same way as the polynomials. They are a generalization of Bézier curves and are advantageous for numerical reasons, as they are locally linearly independent and numerically stable, meaning that small changes in the coefficients result in small changes to the respective spline function and vice versa. The coefficients used as features in the discrimination process resulted from fitting these splines to the formant trajectories using methods provided by the *splines* R package [15]. Figure 3 shows the two spline representations based on the same formant trajectory.

Bent-cable models [4] are an extension of so-called broken stick piecewise-linear models. Linear phases are assumed at the beginning and the end of the trajectory and are approximated by first-degree polynomials which are joined by a interjacent quadratic bend. The full model is given in equation 3.

$$f(x; \beta_0, \beta_1, \beta_2, \tau, \gamma) = \beta_0 + \beta_1 x + \beta_2 q(x; \tau, \gamma)$$

$$q(x; \tau, \gamma) = \frac{(x - \tau + \gamma)^2}{4\gamma} \mathbb{1}\{|x - \tau| \leq \gamma\} + (x - \tau) \mathbb{1}\{x > \tau + \gamma\} \quad (3)$$

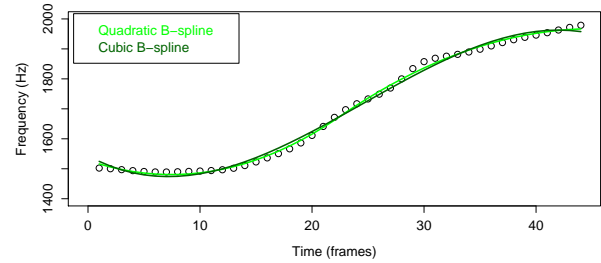


Fig. 3: B-splines fitted to F2 of /æ/ in Kreide

The non-linear transition parameters τ and γ represent the center and half-width of the bend, respectively. These coefficients are used along with the y-intercept, i.e. the absolute formant frequency at the starting point of the trajectory, to characterise the abruptness of the transition in a diphthong. The parameters were estimated using methods provided by the *SiZer* R package [18]. An example is given in figure 4.

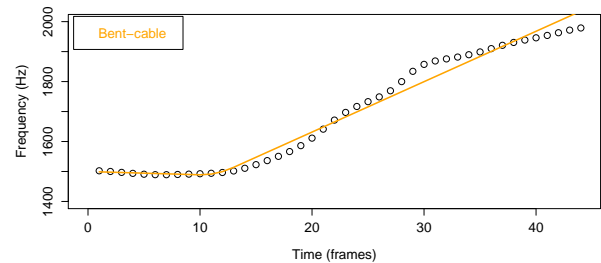


Fig. 4: Bent-cable model based on F2 of /æ/ in Kreide

The rationale behind using this representation was to gain coefficient values that are more easily interpretable with respect to the formant trajectory than those of the aforementioned methods.

3. LIKELIHOOD RATIO FRAMEWORK

The speaker comparison procedure was conducted using the likelihood ratio framework. In this approach, two mutually exclusive hypotheses are considered. The prosecution hypothesis H_0 states that the two sets of measurements on the speech recordings are produced by the same speaker, whereas the defence hypothesis H_1 states that they are produced by different speakers. Using the Bayesian approach of evaluation of evidence, the posterior odds on H_0 given the evidence equal the prior odds

on H_0 , i.e. the ratio of the probabilities of the hypotheses before any speech evidence is used, times the likelihood ratio (see equation 4).

$$\underbrace{\frac{p(H_0 | E)}{p(H_1 | E)}}_{\text{Posterior Odds}} = \underbrace{\frac{p(H_0)}{p(H_1)}}_{\text{Prior Odds}} \cdot \underbrace{\frac{p(E | H_0)}{p(E | H_1)}}_{\text{Likelihood Ratio}} \quad (4)$$

The likelihood ratio represents the probability of observing the evidence, i.e. the difference in feature distributions, assuming the two speakers are the same, versus the probability of this observation if the defence hypothesis H_1 is true. In simple terms, the likelihood ratio is a summary of the additional evidence that is provided by the measurements.

The actual likelihood ratio calculation is carried out using the multi-variate kernel density (MVKD) likelihood ratio formula [1]. The parameters of both suspect and offender samples are each modelled by a multi-variate normal distribution. Two levels of variance are assumed, the within-speaker variability, also assumed normally, and the between-speaker variability, which is modelled by a kernel density estimate based on measurements from a reference population. This analytic formula has been used in several studies that employ formant [17] and f_0 features [6] as well as polynomials and DCT based representations [14, 13].

4. VIENNESE GERMAN DATA

The data used for the evaluation consists of formant trajectories obtained from diphthongs produced by 30 male speakers of Viennese German. They were asked to repeat a sentence containing the word *kreidebleich* ten times, which includes two /æ/ diphthongs, one in primary and one in secondary stressed position, hence resulting in 20 realisations. The diphthongs were hand-labelled and the first three formant tracks were calculated and manually corrected using S.TOOLS-STx [19].

The motivation for using this data was to explore the effects of monophthongisation in the Viennese dialect, a diachronic process which caused the diphthongs /æ/ and /aɔ/ to change into the monophthongs /æ:/ and /ɔ:/ [10]. In the Viennese dialect, monophthongised forms are used, whereas in Standard Viennese German it is a rather gradual process where monophthongised forms are produced mainly in prosodically weak positions [11]. Thus, it is expected that there exists greater variability in the formant trajectories of different speakers, especially for the segment /æ/ in *bleich*.

It must be noted that, since the sentences were recorded in one session, they do not constitute forensically realistic material, as the effects of inter-session variability are neglected. Thus, the study investigates only the general discriminatory potential of the discussed methods.

5. RESULTS

In the performance comparison, cross-validation was used for each method. For each trial, four sets of measurements from each speaker were used. From all 30 speakers, one speaker was selected as offender and one speaker was selected as suspect while the remaining speakers were used for background data in the likelihood ratio calculation. The study limited the number of measurements taken to represent one speaker to ensure the availability of several same-speaker comparisons. Results are presented in the form of detection error trade-off (DET) plots [7] as well as the equal error rate (EER). The log-likelihood ratio cost metric C_{llr} [3] is provided for each method as a measure of the calibration properties of the obtained likelihood ratios.

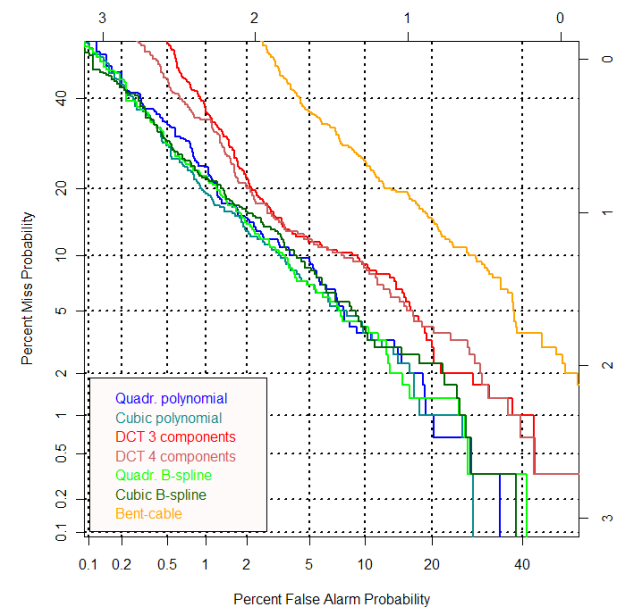


Fig. 5: DET plot comparing the methods applied to raw, non-time-equalised trajectories of F1-3

Figure 5 compares the methods based on error rates achieved in trials performed on the basis of raw, non-interpolated trajectories of the first three formants. As can be seen, the polynomials and B-splines show lower

	EER	C_{llr}
Quadr. polynomial	6.7%	0.2496
Cubic polynomial	6.3%	0.2348
DCT 3 components	9.4%	0.3377
DCT 4 components	9.3%	0.3322
Quadr. B-spline	6%	0.2423
Cubic B-spline	6.3%	0.2584
Bent-cable	17.3%	0.598

Table 1: EER and C_{llr} results based on raw, non-time-equalised trajectories of F1-3

error rates compared to the DCT-based methods and the bent-cable method, which displays substantially higher error rates. Table 1 gives an outline of the results in terms of EER and C_{llr} .

Previous studies [12, 14] showed differences in performance of parametric representations when applying time-normalisation to the formant trajectories. Figure 6 compares the methods based on interpolated formant tracks.

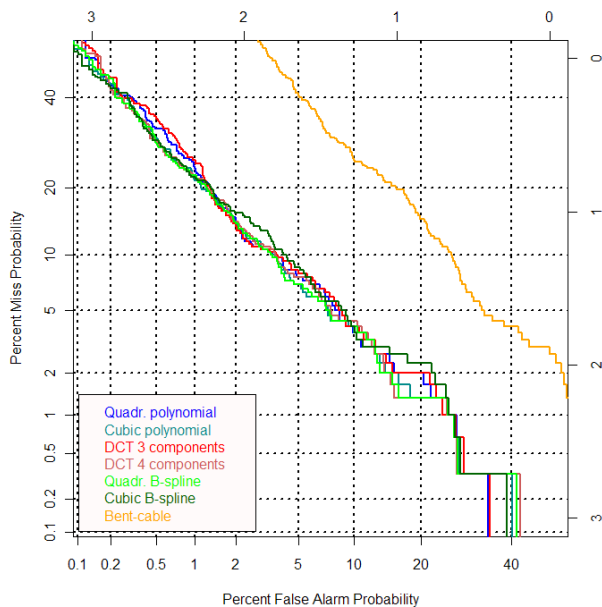


Fig. 6: DET plot comparing the methods applied to time-equalised trajectories of F1-3

The polynomial and B-spline representations as well as the bent-cable method basically show the same error

	EER	C_{llr}
Quadr. polynomial	6.7%	0.2519
Cubic polynomial	6%	0.2435
DCT 3 components	6.7%	0.2558
DCT 4 components	6.6%	0.2461
Quadr. B-spline	6%	0.2424
Cubic B-spline	6.3%	0.2586
Bent-cable	17.8%	0.6057

Table 2: EER and C_{llr} results based on time-equalised trajectories of F1-3

rates, but there is quite an improvement for the DCT-based approaches. Table 2 summarises these results.

These trials were performed by using four measurements of each speaker and arranging them to contain two sets derived from /aɛ/ in *kreide* and two in *bleich*, which resembles a rather optimistic setting. Another set of trials was performed in which four sets contain only measurements in one context and one set contains two of both (to utilise the total set of 20 measurements) to simulate the worst-case in terms of feature composition. Figure 7 shows the results of trials separated for their phonetic context.

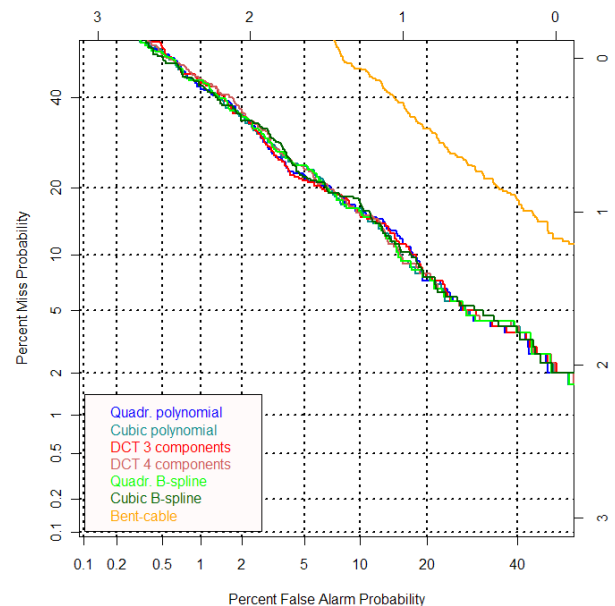


Fig. 7: DET plot comparing the methods applied to time-equalised trajectories of F1-3 without mixing of context

	EER	C_{llr}
Quadr. polynomial	13.4%	1.009
Cubic polynomial	13%	1.0191
DCT 3 components	13.6%	1.007
DCT 4 components	13.1%	1.0231
Quadr. B-spline	12.9%	1.0193
Cubic B-spline	12.9%	1.0561
Bent-cable	26%	1.0185

Table 3: EER and C_{llr} results based on time-equalised trajectories of F1-3 without mixing of context

As can be seen, the performance of all methods decreases severely, with B-splines achieving the lowest EER of 12.9% (see table 3). This signals that there is considerable variation between the coefficients derived from the two segments.

Another condition that is relevant in a forensic setting is presented when recordings are taken from telephone conversations. In these cases, the bandpass characteristics of the telephone channel can cause corruption of the first formant. For this reason, comparisons were made based on trials that utilise only coefficients derived from the trajectories of the second and third formant.

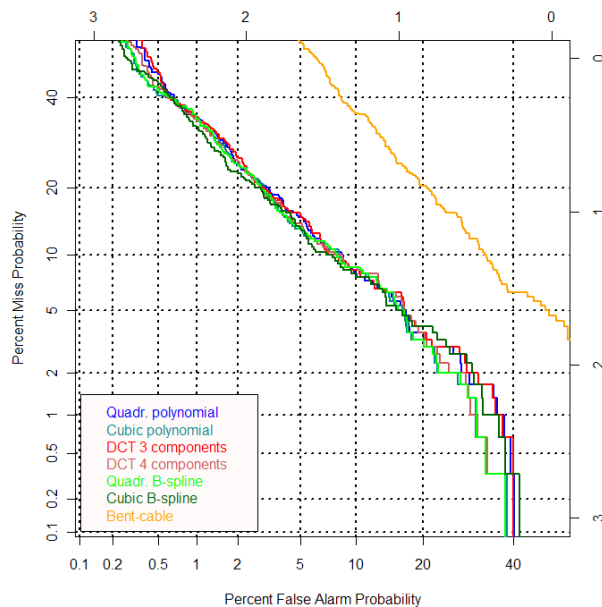


Fig. 8: DET plot comparing the methods applied to time-equalised trajectories of F2 and F3

	EER	C_{llr}
Quadr. polynomial	9%	0.355
Cubic polynomial	8.8%	0.3339
DCT 3 components	9%	0.3579
DCT 4 components	9%	0.3374
Quadr. B-spline	8.9%	0.3341
Cubic B-spline	8.7%	0.3389
Bent-cable	20.3%	0.6986

Table 4: EER and C_{llr} results based on time-equalised trajectories of F2 and F3

The results of these trials are depicted in figure 8. The polynomials, B-splines and DCT-based methods cluster together at EER values of close to 9%, while the bent-cable model displays a much higher error. EER and C_{llr} values are given in table 4.

6. CONCLUSIONS

The study evaluated the performance of different parametric representations of formant trajectories. In general, polynomials and B-splines consistently displayed low error rates, as do DCT-based methods with the exception of trials based on non-time-equalised formant tracks.

Bent-cable coefficients, however, show substantially higher error rates than the other methods. This can partly be attributed to the automatic fitting of the models which is sensitive to the presence of additional bends in the trajectory, e.g. at the end of the segment due to coarticulation, which are in turn modelled by the quadratic bend, thus not characterising the actual transition in the diphthong.

Results achieved in tests where segments with different context were kept separate suggest that the general performance is highly dependent on the composition of the input data sample with respect to its phonetic context and prosodic position. This can be mainly ascribed to two aspects. First, there is an expected increase in variability between the two segments due to the aforementioned Viennese monophthongisation process which induces additional variability due to /aɛ/ in *bleich* being more susceptible to monophthongisation because of its secondary stressed position. The second reason that must be considered is the rather artificial speech situation in which the recordings were made. For this reason, additional tests should be conducted based on actual forensic data.

7. REFERENCES

- [1] Colin G. G. Aitken and David Lucy. Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53(1):109–122, 2004.
- [2] Timo Becker, Michael Jessen, and Catalin Grigoras. Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models. In *Proceedings of Interspeech 2008 incorporating SST'08*, pages 1505–1508, Brisbane, September 2008. ISCA.
- [3] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20:230–275, 2006.
- [4] Grace Chiu, Richard Lockhart, and Richard Routledge. Bent-cable regression theory and applications. *Journal of the American Statistical Association*, 101(474):542–553, 2006.
- [5] Catalin Grigoras, Michael Jessen, and Timo Becker. Forensic speaker verification using long term formant distributions and likelihood ratios. In *50th European Academy of Forensic Sciences Conference*, Glasgow, September 2009.
- [6] Yuko Kinoshita, Shunichi Ishihara, and Phil Rose. Beyond the long-term mean: Multivariate likelihood ratio-based FSR using F0 distribution parameters. In *Proceedings of the IAFPA*, page 15, 2007.
- [7] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech '97*, pages 1895–1898, 1997.
- [8] Kirsty McDougall. *The Role of Formant Dynamics in Determining Speaker Identity*. PhD thesis, Department of Linguistics, University of Cambridge, 2005.
- [9] Kirsty McDougall and Francis Nolan. Discrimination of speakers using the formant dynamics of /u:/ in British English. In J. Trouvain and W. Barry, editors, *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 1825–1828, Saarbrücken, August 2007. ICPhS.
- [10] Sylvia Moosmüller. The Process of Monophthongization in Austria (Reading Material and Spontaneous Speech). *Papers and Studies in Contrastive Linguistics*, 34:9–25, 1998.
- [11] Sylvia Moosmüller and Ralf Vollmann. 'Natürliches Driften' im Lautwandel: die Monophthongierung im österreichischen Deutsch. *Zeitschrift für Sprachwissenschaft*, 20/1:42–65, 2001.
- [12] Geoffrey Stewart Morrison. Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *International Journal of Speech, Language, and the Law*, 15(2):249–266, 2008.
- [13] Geoffrey Stewart Morrison. Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125(4):2387–2397, April 2009.
- [14] Geoffrey Stewart Morrison and Yuko Kinoshita. Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories. In *Proceedings of Interspeech 2008 incorporating SST'08*, pages 1501–1504, 2008.
- [15] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.
- [16] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.
- [17] Phil Rose, Yuko Kinoshita, and Tony Alderman. Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. In *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*, pages 329–334, 2006.
- [18] Derek Sonderegger. *SiZer: SiZer: Significant Zero Crossings*, 2008. R package version 0.1-0.
- [19] STx, 2010. <http://www.kfs.oeaw.ac.at/>.