Parametric Diphthong Formant Trajectory Representations for Forensic Speaker Recognition

Ewald Enzinger[1]

[1] *Acoustics Research Institute, Austrian Academy of Sciences, Email: ewald.enzinger@oeaw.ac.at*

Introduction

Acoustic phonetic features used in a forensic setting are most commonly static, either involving formant measurements at phonetic targets or statistic estimates of parameters over parts or the whole speech sample, e.g. long term formant distributions [3]. Recently, interest has increased in parameters that capture temporal-dynamic properties of speech segments. The proposed methods are based on deriving parametric representations of formant trajectories to utilise the resulting coefficients as features in the speaker identification process.

Parametric and instantaneous formant trajectory representations

The evaluation presented in this paper adopts three methods for deriving parametric representations from formant trajectories. *Polynomial curves* fitted to formant trajectories [8] and coefficients of *discrete cosine transform (DCT)* [9] were used in addition to *B-spline* representations, i.e. pairwise polynomials fitted to the formant trajectories. Figure 1 compares the resulting functions for a realisation of /aɛ/ (see below).

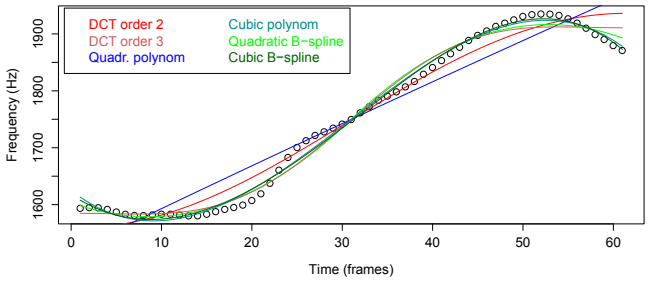

Figure 1: Comparison of parametric curves fitted to F2 of /aɛ/ in *Kreide*

B-splines are a generalization of Bézier curves. They 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.

Error rates achieved using these representations were compared with those based on formant measurements taken at presumed phonetic targets at 20% and 80% of the segment duration (*Simulated dual-target*), as well as using 9 formant measurements at *10% time intervals* [5].

Likelihood ratio calculation

The different features were entered into a multi-variate kernel density (MVKD) likelihood ratio formula [2]. 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. Both are estimated from the background data. This analytic formula has been used in several studies that employ formant [10] and f0 features [4] as well as polynomials and DCT based representations [8, 9].

Viennese German data

The data used for the evaluation consists of formant trajectories obtained from diphthongs produced by 30 male speakers of Viennese German. The two segments /aɛ/ were taken from the word *kreidebleich*. They were hand-labelled and the first three formant tracks were calculated and manually corrected using S_TOOLs-STx [1].

The segments have been selected to explore the effects of monophthongisation in Viennese German, a diachronic process which caused the Standard Austrian German diphthongs /aɛ/ and /aɔ/ to change into the monophthongs /æː/ or /ɛː/ and /ɒː/ or /ɔː/ in the Viennese Dialect [6]. 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 [7], such as /aɛ/ in *bleich*.

Results

For performance comparison, cross-validation was used for each method in the evaluation. 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. Figure 2 compares the performance of the methods. The lowest error rates were achieved by quadratic B-splines (EER 6%) and cubic polynomials (EER 6.3%). The simulated dual-target and 10% interval methods showed higher error rates, achieving 7.4% and 8% EER, respectively. The methods based on DCT coefficients displayed the highest error rates (EER 9.3).

In these trials, the four measurements of each speaker were arranged to contain two sets derived from /aɛ/ in *kreide* and two in *bleich*, which resembles a rather op-

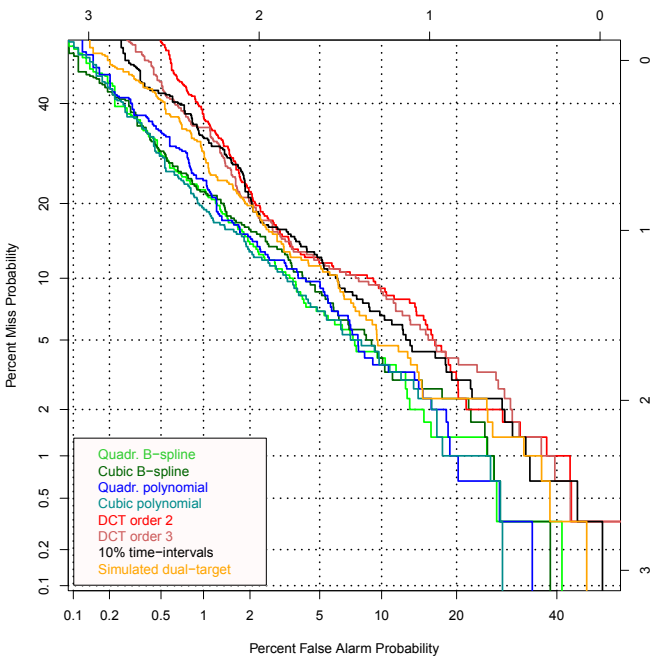

Figure 2: DET plot comparing the performance of the methods based on shuffled trial data sets

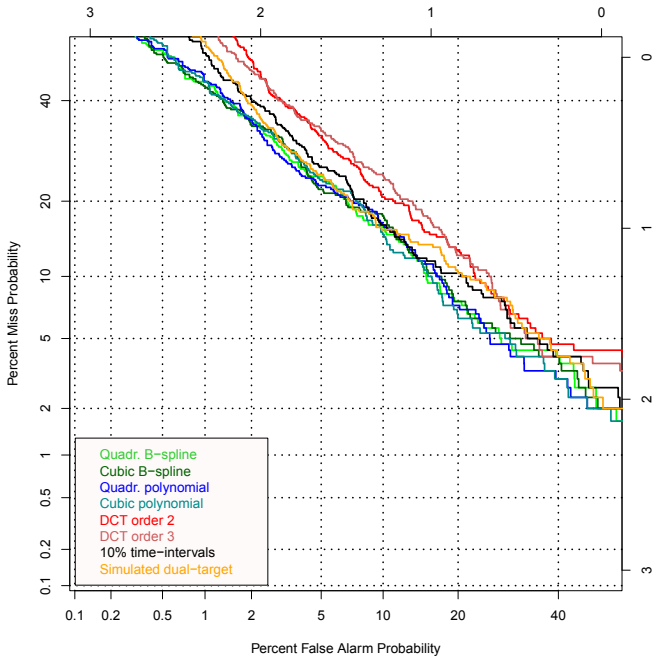

Figure 3: DET plot comparing the performance of the methods based on separated trial data sets

timistic 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). Figure 3 shows the results of trials separated for contexts. As can be seen, the performance of all methods decreases severely, with cubic polynomials achieving the best error rate of 12.3% EER.

Conclusions

The evaluation compared the performance of different parametric representations of formant trajectories. In general, polynomials and B-splines outperformed methods based on DCT coefficients as well as instantaneous measurements. However, the general performance is highly dependent on the composition of the input data sample with respect to it's phonetic context and prosodic position. The Viennese monophthongisation process induces additional variability due to /aɛ/ in *bleich* being more susceptible to monophthongisation because of it's secondary stressed position.

References

[1] STx, 2010. `http://www.kfs.oeaw.ac.at/`.

[2] Colin G. G. Aitken and David Lucy. Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53(1):109–122, 2004.

[3] 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.

[4] 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.

[5] Kirsty McDougall. *The Role of Formant Dynamics in Determining Speaker Identity*. PhD thesis, Department of Linguistics, University of Cambridge, 2005.

[6] Sylvia Moosmüller. The Process of Monophthongization in Austria (Reading Material and Spontaneous Speech). *Papers and Studies in Contrastive Linguistics*, 34:9–25, 1998.

[7] Sylvia Moosmüller and Ralf Vollmann. 'Natürliches Driften' im Lautwandel: die Monophthongierung im österreichischen Deutsch. *Zeitschrift für Sprachwissenschaft*, 20/1:42–65, 2001.

[8] 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.

[9] 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.

[10] 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.