Formant tracking in Sound Tools eXtended (STx) 5.0

Anton Noll, Michael Pucher, Carina Lozo

Acoustics Research Institute (ARI), ÖAW, Vienna, Austria, Email: {anton.noll, michael.pucher, carina.lozo}@oeaw.ac.at

Abstract

In this paper, we introduce the new formant tracker of Sound Tools eXtended (STx) version 5.0, an acoustic speech and sound processing application. Formants occur at the resonant frequencies of the human vocal tract and can be correlated with articulatory features such as tongue height and backness. They are indispensable for acoustic phonetic analysis since they allow for an identification of vowel sounds. Our formant tracker uses dynamic programming to find the best path among formant candidates derived from Linear Predictive Coding (LPC) analysis. The transition and observation costs in the search algorithm are optimized with manually corrected formant tracks. The new version of STx 5.0 also has an interface to WebMAUS that allows automatic phonetic transcription and segmentation of speech for different languages. Furthermore STx 5.0 contains an integrated, simplified and compact GUI, designed for speech analysis for phoneticians, linguists, psychologists, and researchers in related fields.

Introduction

Sound Tools Extended (STx) is an acoustic speech and signal processing application for Windows. It provides tools to analyse, visualise, segment, and annotate wave files. For non-commercial, scientific and educational purposes, we offer STx free for download [2].

STx has an adjustable formant tracker and allows for manual correction and reassignment of formant trajectories. Information on segments and formants is stored in an automatically generated text file that can be used for further processing and analysis. The settings for all shown functions are readjustable in a window.

It uses its own annotation file format, but can import and export PRAAT TextGrid files [3]. The new version of STx 5.0 also has an interface to WebMAUS that allows automatic phonetic transcription [4]. Further features are automatic detection of silences/pauses, a speech recorder, real time spectrogram, an anonymization tool, and an easy editing of sound files.

STx has been used in Phonetics, Speech Synthesis, Forensics Bioacoustics, Analysis of noise, Musicology, and Neuroscience. A more detailed description of STx and its features can be found in [1].

Formant tracking review

The main application of formant tracking nowadays lies in acoustic phonetics and forensic speech science. Different approaches have been applied for formant tracking using Hidden Markov Models (HMM) [5], graphical models [6], or dynamic programming [7]. It has also been shown that the inclusion of context-dependent phonemic information can improve formant tracking results [8]. In the past formant tracking was also important for speech synthesis [10], which has changed as other synthesis paradigms have been developed in the last decades [11]. The most prominent approach for detection of formant candidates is Linear Prediction (LP) [9].

Training data

As training material we use the first 4 formants of our corpus "Österreichisches Deutsch (OeD)" ("Austrian German"), a corpus that was collected within different projects between 2008 and 2019. The subset consists of 250 recordings of male and female speakers with 30000 formant data sets where most are manually verified and/or corrected. All formant tracks have comparable analysis settings with a frame length of $\approx$ 40ms and a hop size of $\approx$ 2.5ms. LP analysis is used to estimate formant candidates.

Our guiding principle for selecting formant tracks for training was that ideally segments should be syllables, words, or phrases and should contain vowel transitions to also have these transitions included in the model. This principle led to the selection of material where

- we only use formant tracks with at least 40 measurements to exclude random or spurious measurements.
- We only use segments with a length between 40-400ms. The lower limit comes from the frame length, the upper border is due to the fact that tracks which were corrected by phoneticians are mostly short ones.
- Only select formant tracks computed with frame length of 40ms – 50ms and hop size 2ms – 2.5ms.

This leads to 30000 formants from 250 recordings. An analysis of different subsets of this corpus showed that the model parameters have similar distributions, which allows use to conclude that the data is homogeneous. The number of male and female speakers is in a similar range. The speaker database also contains additional information on gender, age, dialect/standard language variety, and word lists for recordings in Standard German. What we did not evaluate at the moment is if the selection is phonetically balanced.

Algorithm1: Formant tracking with missing formants

The formant tracks are computed using the Viterbi algorithm, a dynamic programming method shown in Figure 1. $\delta_j(t)$ is the probability of being in state j at time t that is computed by taking the maximum probability of all previous states and multiplying it with the transition probability a_{ij} and observation probability $b_j(o_t)$:

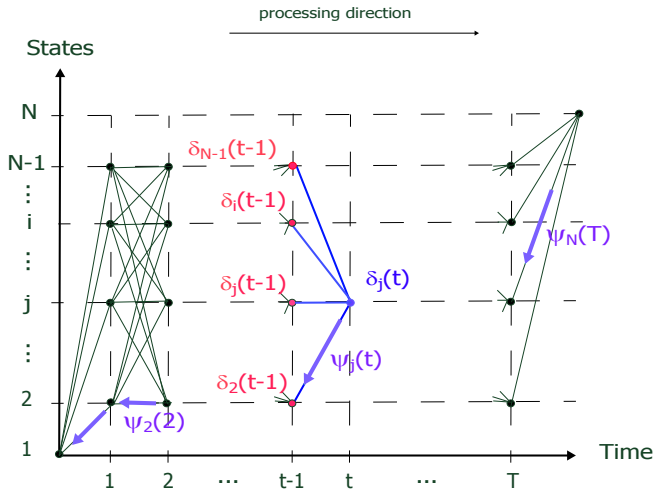

Figure 1: Viterbi algorithm.

$$\delta_j(t) = \max_{1<i<N} [\delta_i(t-1)a_{ij}]\, b_j(o_t).$$

Together with $\delta_j(t)$ we also have to keep track of the state that maximized the probability at each step with the function $\psi_j(t)$:

$$\psi_j(t) = \operatorname*{argmax}_{1<i<N} [\delta_i(t-1)a_{ij}]\,.$$

After computing $\delta_j(t)$ and $\psi_j(t)$ for all times t and all states N, we can use $\psi_j(t)$ for backtracking to find the best state path in the model as shown in Figure 2.

In our case a state is defined as a possible combination of formants, a formant pattern e.g. {F1, F2}, {F1, F3}, {F1, F4}, {F1, F2, F3}, ... This state space also takes into account that not all formants are present all the time. With 4 formants there are $2^4 - 1 = 15$ different possible formant patterns, the number of subsets of a set with 4 elements minus the empty set that we are not considering. The transition probability between the empty set and other formant patterns could be easily computed from the database, but the computation of the observation probability is less straightforward. We therefore use F0 tracking to decide if formants should be present or not and then do the formant tracking.

The transition probabilities a_{ij}, giving the probablity of going from one state j at time $t-1$ to another state k at time t is given by:

$$a_{ij} = P(S_t = s_k | S_{t-1} = s_j)$$

and estimated from the formant training data.

For each frame a set of formant frequency candidates is computed, using the LPC speech model. We use only frequencies because our training data set does not contain band-width values. For these candidates the possible states (per frame) and transitions (frame to frame) are build. The state/transition probabilities are used to compute the probabilities of possible paths and a backtracking algorithm is then applied to get the best path.

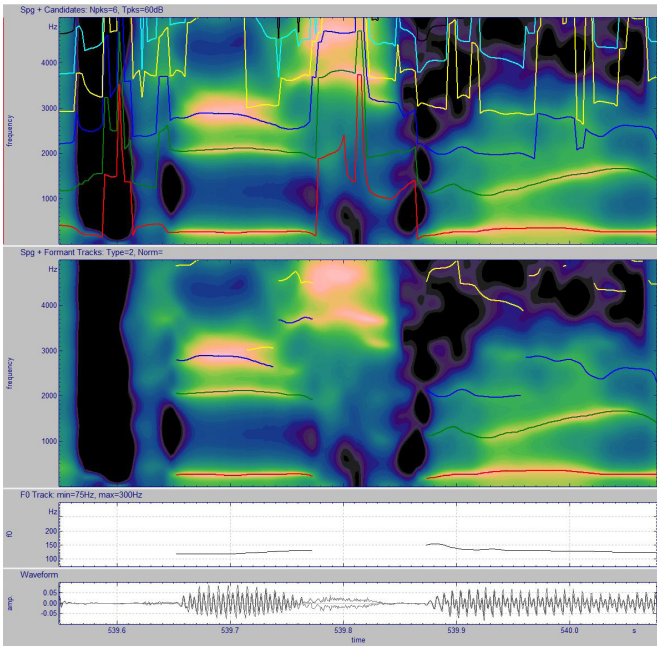

Figure 2: Formant tracker showing the formant candidates (top) and formant tracks after Viterbi search (bottom) for the word "diesmal" [d i: s m a: l] of a male speaker.

For implementing the observation probability $b_j(o_t)$, where o_t is a number of formant candidates and j is a formant pattern we use formant ranges Fmin$(1-n)$ and Fmax$(1-n)$ to associate a formant pattern with formant candidates and compute the cost of the formant pattern / formant candidate combination. The formant ranges are defined as F1=50 ... 1550, F2=400 ... 3400, F3=1400 ... 4600, F4=2500 ... 6000 (Hz). Furthermore we also use the probability of a certain formant pattern e.g. $P(\{F1, F2\})$ that can be estimated from the training data.

Algorithm2: Formant tracking with formant sub-ranges

The second variant takes into account that the first variant produces the same $\delta_i(t-1)a_{ij}$ values at a certain time for different states i. In this case the best transition is decided with the mean distance. Since vowels are not only discerned by the absolute value of the formants but also by the relation between the formants we applied the following model.

Each formant is defined by a generously chosen formant range. This range is then divided into multiple bands (e.g. low / medium / high). The states are then defined by all formant and band combinations, e.g. {(F1, low), (F2, high), (F3, medium), (F4, low)}, ... without considering missing formants. This leads to $3^4 = 81$ possible combinations. Also taking missing formants into account would generate $4^4 - 1 = 255$ possible combinations. The algorithm proceeds in the same way as in Algorithm1.

Conclusion

We have presented the new formant tracker of Sound Tools eXtended (STx) version 5.0. At the moment Algorithm1 is implemented as a script that can be run from STx, and Algorithm2 is available in a test version. The algorithms are based on the Viterbi algorithm and use estimated parameters from a formant database. The algorithms works with a small amount of training data and also allow for the manual tuning of several of it's parameters like formant ranges. Our algorithms are similar to the formant tracker proposed in [7] but we also include missing formants in our state space. Its integration into the powerful STx program will make it usable for researchers in many different fields.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) project DiÖ - Deutsch in Österreich (F6002-G23).

References

[1] A. Noll, J. Stuefer, N. Klingler, H. Leykum, C. Lozo, J. Luttenberger, M. Pucher, C. Schmid, *Sound Tools eXtended (STx) 5.0 – a powerful sound analysis tool optimized for speech*. In Proccedings of Interspeech 2019 - Show&Tell, Graz, Austria, 2019.

[2] Sound Tools Extended (STx) - Intelligent Sound Processing, `https://www.kfs.oeaw.ac.at/stx`

[3] P. Boersma, D. Weenink, *Praat: doing phonetics by computer*, Amsterdam, 2017. [Online]. Available: `http://www.praat.org`

[4] F. Schiel, *Automatic phonetic transcription of non-prompted speech*, in Proceedings of the ICPhS, 1999, pp. 607–610.

[5] G. Kopec, *Formant tracking using hidden Markov models and vector quantization*, in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, no. 4, pp. 709-729, August 1986.

[6] J. Malkin, Xiao Li and J. Bilmes, *A graphical model for formant tracking*, Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Philadelphia, PA, 2005, pp. I/913-I/916 Vol. 1.

[7] K. Xia, C. Espy-Wilson, *A new strategy of formant tracking based on dynamic programming*, In ICSLP-2000, vol.3, 55-58.

[8] M. Lee, J. van Santen, B. Mobius and J. Olive, *Formant tracking using context-dependent phonemic information*, in IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 741-750, Sept. 2005.

[9] R. C. Snell and F. Milinazzo, *Formant location from LPC analysis data*, in IEEE Transactions on Speech and Audio Processing, vol. 1, no. 2, pp. 129-134, April 1993.

[10] A. Acero, *Formant analysis and synthesis using hidden Markov models*, In EUROSPEECH'99, 1047-1050.

[11] C. Lozo, J. Luttenberger, M. Pucher, *The thought collective behind thirty years of progress in speech synthesis*. Proceedings of the 3rd International Workshop on the History of Speech Communication Research, Vienna, Austria, 2019. Studientexte zur Sprachkommunikation, TUDpress, Band 94, pp. 49-58.

[12] Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, I. Andrianakis, *Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing*, Computer Speech & Language, Volume 21, Issue 3, 2007, Pages 543-561.