Evaluation of Automatic Formant Trackers

F. Schiel, Th. Zitzelsberger
Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität
Schellingstr. 3, 80799 München, Germany
{schiel,tzom}@bas.uni-muenchen.de

Abstract

Four open source formant trackers, three LPC-based and one based on Deep Learning, were evaluated on the same American English data set VTR-TIMIT. Test data were time-synchronized to avoid differences due to different unvoiced/voiced detection strategies. Default output values of trackers (e.g. producing 500Hz for the first formant, 1500Hz for the second etc.) were filtered from the evaluation data to avoid biased results. Evaluations were performed on the total recording and on three American English vowels [i:], [u] and [ʌ] separately. The obtained quality measures show that all three LPC-based trackers have comparable RSME error results that are about 2 times the inter-labeller error of human labellers. Tracker results are biased considerably (consistently too high or low), when the parameter settings of the tracker are not matched to the speaker's sex. Deep Learning appears to outperform LPC-based trackers in general, but not in vowels. Deep Learning has the disadvantage that it requires annotated training material from the same speech domain as the target speech, and a trained Deep Learning tracker is therefore not applicable to other languages.

Keywords: formant, tracker, evaluation, VTR-TIMIT

1. Introduction

This paper presents the methodology and results of a technical evaluation of four open source algorithms for automatic formant tracking (in the following referred to as 'formant trackers'). The motivation for such a study is obvious: currently a small number of open source formant trackers is widely used by speech scientists and speech engineers, but to our knowledge there exists no objective evaluation of the quality of more than two algorithms based on the same data set (Deng et al, 2006 evaluate their own formant tracker compared to WaveSurfer based on VTR-TIMIT). This is due partly to the fact that a manually controlled refence corpus for formant tracks is difficult to obtain. Fortunately, with the publication of the VTR-TIMIT (Vocal Tract Resonances TIMIT) corpus by Deng et al (2006) we are now able to perform such an evaluation for US American English.

2. Formants

A formant is a resonance in the speech signal caused by the geometry of the physiological tubular system of the speaker's vocal tract. Formants are considered to be the primary phonetic feature for distinguishing vowel classes as well as place of articulation in consonant-vowel transitions. Furthermore, since formants are determined by the ideosyncratic physiological form of a speaker's vocal tract, they play a crucial role in forensic speaker recognition, automatic speaker identification and verification, sex and age recognition (e.g. Rose, 2003, pp. 221).

A formant is typically defined by three parameters: the center frequency (often called formant frequency), the bandwidth and the amplitude of the resonance. Technically, formants in a digitized speech signal are often decribed as complementary poles in the z-transform of the vocal tract filter, where the radial position of the pole defines the center frequency and the distance to the unit circle (and distance to neigboring poles/zeros) defines bandwidth and amplitude.

The lower formants 1-5 are widely used as phonetic features in linguistic-phonetic and forensic analysis but also as basic features in speech technology applications (such as speech morphing, speech and speaker recognition). It is therefore not surprising that the development of algorithms to detect and track lower formants automatically in a speech recording has a long tradition going back to Rabiner & Schafer (1970).

3. Formant Trackers

Since formants are basically caused by an acoustical filter operation where the glottal source signal is filtered by a infinite-impulse-response filter, i.e. a filter having 5 or more complementary poles in its z-transform, they cannot be determined analytically from the recorded speech signal without prior knowledge of the source signal (which is usually not available). Most algorithms to estimate formant parameters from the recorded speech signal therefore either apply homo-morphic analysis of the spectral envelope (e.g. cepstral analysis followed by a peak-picking strategy), or LPC analysis (Markel & Grey, 1982) to estimate the z-transform followed by a pole-picking strategy, or trained pattern recognition techniques (e.g. support vector machine, random forest or deep learning). The latter requires a training set of labelled formant tracks and is in most cases language dependent, while the former two approaches are inherently language independent and do not require any training material.

The task of formant tracking is further complicated by the fact that some algorithms assume a voiced signal for analysis, i.e. the spectral envelope encloses a harmonic spectrum consisting of a fundamental frequency line and the respective harmonic spectral lines at multiples of the fundamental frequency. Such algorithms typically produce more or less random results when applied to unvoiced parts of the speech signal; therefore formant track algorithms are often combined with a voiced-unvoiced detector to suppress formant analysis in unvoiced parts of the speech signal. Since the voiced-unvoiced detection in itself is error prone (e.g. in creaky voice), the output of the combined algorithm is influenced by the performance of both algorithms.

In this study four formant trackers are investigated:

- **PRAAT**, the built-in formant tracker of the praat tool by Boersma & Weenink (2017), 'Burg' method (cf. Childers 1978, pp. 252)

- **SNACK**, the formant tracker (version 2.2) of the Snack Sound Toolkit of KTH Stockholm by Kåre Sjölander (2017)
- **ASSP**, the formant tracker *forest* (version 2.8) contained in the Advanced Speech Signal Processor library by M. Scheffer (Scheffer, 2017), also contained in the R language package *wrassp*, and part of the Emu database management system (EMU-SMDS, Winkelmann 2017)
- **DEEP** : DeepFormant, a formant tracker (Keshet, 2017) based on deep learning techniques and trained on the training set of VTR-TIMIT (Dissen & Keshet, 2016).

The first three formant trackers are based on LPC analysis; no homo-morphic formant tracker was evaluated in this study (cf. Kammoun et al, 2006 for a discussion of LPC vs. homo-morphic formant analysis).

4. Test Data VTR-TIMIT

Vocal Tract Resonance TIMIT (VTR-TIMIT) is an open source subcorpus annotation of TIMIT[1] (Garofolo et al, 1992) with 516 manually annotated recordings spoken by 186 (113m and 73f) speakers of American English. The subcorpus contains phonetically compact and phonetically rich sentences, and no dialectal speech. The speech was first analysed by the algorithm described in Deng et al, 2004, and subsequently hand-corrected. The manual correction was performed by a group of labelers based on visual inspection of the first three formants in the spectrogram (higher formants, bandwidths and amplitudes were not corrected). Inter-labeller agreement tests on a small sub-sample (16 sentences per 5 different labeller-pairings) yielded average frequency deviations of about 78Hz for the first formant (F1), 100Hz for F2 and 111Hz for F3 (Deng et al, 2006).

For technical reasons the VTR-TIMIT formant reference tracks are continuous over the total recording, i.e. there is no indication of where the speech is voiced or where formants are or are not visible in the spectrogram. Formants in unvoiced or silent parts of the signal are either interpolated linearly from the two adjacent voiced parts or horizontally extended at the initial or final voiced portion of the recording. This interpolation facilitates an evaluation of formant tracker output independently of the voiced-unvoiced detection of the tracker algorithm (because for every time frame of the recording there exists a reference value); on the other hand the resulting quality measure might be compromised: if a tracker is 'conservative' in the sense that it produces output only for the parts of the input signal where it is quite confident (clearly voiced parts), then this tracker will outperform other trackers who produce results in parts of the signal where the tracking is compromised for instance by creaky voice or noise associated with consonantal constrictions etc.

Since the tracker DeepFormant was trained on parts of VTR-TIMIT, the following evaluations of DeepFormant were only performed on the test part of VTR-TIMIT (a subset with 8f and 16m speakers).

[1] TIMIT itself and thus the signals of VTR-TIMIT are not open source; refer to the Linguistic Data Consortium.

5. Evaluation Methodology

Quality Measures

Two quality measures were calculated to quantify the distance of formant tracker outputs to the annotation reference:

- root mean squared error (**RMSE**): to quantify overall performance (zero being perfect match)
- average error (**AVG**) to indicate systematic errors: positive values indicate that the tracker tends to calculate formant estimates in average too low; negative values indicate formant estimates are too high; zero indicate perfect symmetry of errors around the annotation reference.

Histograms of the AVG errors were plotted to identify multi-modal distributions, for instance caused by systematic formant confusion errors.

Formant Tracker Parameters

The four formant trackers were evaluated with their default parameter settings, except for the frame step size which was set to 10msec and the window length which was set to 25msec for all trackers to yield comparable number of evaluation frames. Other parameters were not changed under the assumption that the developpers have choosen these default parameter sets to optimize for best performance. As we will see in the results, using the default settings causes differences in measurement quality depending on the speaker's sex, since two of the trackers, SNACK and ASSP, use default parameters optimized for male speakers, while PRAAT uses a parameter set optimized for female speakers. However, using speaker-individual parameter settings in the evaluation pose a problem, since not all trackers offer a standard parameter set for female and male speakers. Instead we propose to inspect the results of trackers sorted according to their default sex parameter setting, i.e. to look at the results of female speakers for PRAAT and the results of male speakers for SNACK and ASSP.

The tracker DeepFormant has no parameters to influence speaker sex, but is trained to a dominantly male training data set (67f vs. 95m speakers) and is therefore expected to perform slightly better on male speech. Table 1 lists the available parameters and chosen settings (defaults are marked with asterix *).

Parameter	ASSP	SNACK	PRAAT	DEEP
formants	4*	4*	5*	4*
LPC	18*	12*	10*	n/a
preemph	0.96*	0.7*	50Hz/oct*	unknown
window	blackman*	cos4*	gauss*	unknown
w. length	25ms	25ms	25ms*	unknown
stepsize	10ms	10ms	10ms	10ms*

Table 1: formant tracker parameter sets

Tests

As mentioned earlier formant trackers have different strategies to distinguish between parts of the signal where formants can be detected versus parts of the signal that are unvoiced (or otherwise compromised in a way that no formants can be detected). To prevent quality measures

from being skewed by this behavior we decided to run several different tests:

1. **NORM**: all trackers use the same voiced-unvoiced detection. We used the pitch detector in praat (which is an independent tool from the praat formant tracker) to determine in all test recordings when tracker output is to be evaluated. If a tracker does not deliver 'real' results within these defined areas, the results were excluded. For instance, SNACK will output default formant values in parts of the signal. Fortunately, these can easily be detected and filtered from the evaluation data (less than 6% loss of evaluation data).
2. **DEFAULT**: all trackers decide individually for which portions of the recording formant values are produced. Again, detectable default values are filtered from the evaluation. This is basically the 'normal' way to use a formant tracker, since it is very unlikely that a user will not accept the built-in voiced-unvoiced decision of a formant tracker.
3. **VOWELS**: tracker outputs were restricted to the same voiced-unvoiced segmentation as in test NORM, but additionally restricted to the segments of the three vowels [i:], [u] and [ʌ] which are roughly the corner positions in the American English vowel space. This test was motivated by the fact that the first three formants are the predominant features for vowel quality.

6. Results

For brevity only RMSE results are presented in detail; AVG results are given in the text and can also be seen in the example histograms.

6.1 Test Condition NORM

Table 2 lists the RMSE measures in Hz in test condition NORM for formants F1...F3 and for female (f) and male (m) speakers. As expected, absolute errors increase with formant order; the relative errors are in about the same range for all three formants. Errors of the DEEP tracker are exceptionally low, but the comparison with the remaining three trackers is not fair since the test set was much smaller for DEEP (see section 4) and DEEP was trained to VTR-TIMIT and therefore has an advantage. Tracker PRAAT has problems with male speakers, while SNACK and ASSP perform better on males. However, comparing the average RMSE error value for female speakers in PRAAT (194Hz) with male speakers in SNACK/ASSP (234/177Hz) we can see that the performance is in the same range (underlined values in Table 1). It is unlikely that these differences between trackers are significant, considering that averaged human inter-labeller errors are in the range of 96Hz (see section 4). But it is clear that the LPC-based trackers in average (210Hz error) perform significantly worse than human labellers (about 2 times worse).

	F1		F2		F3	
	f	**m**	**f**	**m**	**f**	**m**
SNACK	126	**100**	291	**227**	313	**375**
ASSP	113	**96**	479	**211**	512	**225**
PRAAT	**116**	234	**217**	338	**249**	404
DEEP	120	97	195	167	252	169

Table 2: RMSE errors in the test condition NORM

Looking at averaged AVG measures, the main result of the evaluation is that ASSP tends underestimate formant frequencies F2/F3 for female speakers (200Hz too low), while PRAAT does the opposite, i.e. it overestimates all three formant frequencies for male speakers (140Hz too high). The averaged AVG errors for SNACK and DEEP are quite balanced.

Selected AVG Histograms

Fig. 1 shows the AVG error histogram of F3 estimated by SNACK. Male values (pink) display a second peak at about -1000Hz AVG error (= estimated 1000Hz too high), indicating that SNACK in some cases confuses F4 with F3 (tracks F3 in the location of F4); the same error is not visible for female speakers (blue), probably because female F4 is much higher than male F4 and therefore not easily confused with F3.

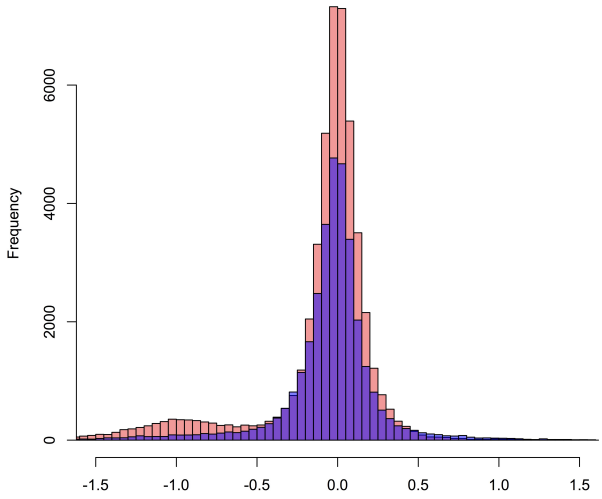

Figure 1: AVG error histogram of F3 estimated by SNACK; positive AVG error values denote estimates lower than reference, and vice versa; female speakers are light blue, male speakers are pink, overlap dark blue.

Fig. 2 shows the histogram of AVG errors for F2 estimated by ASSP. Here female speakers (light blue) show more positive AVG errors, indicating that the ASSP estimates for F2 for female speakers are often too low. In contrast to Figure 1 there is no visible second peak which means that these errors are probably not caused by classical formant confusion.

6.2 Test Condition DEFAULT

The RMSE measures obtained in test condition DEFAULT were almost as in the test condition NORM (for brevity we do not show the table in this abstract). The

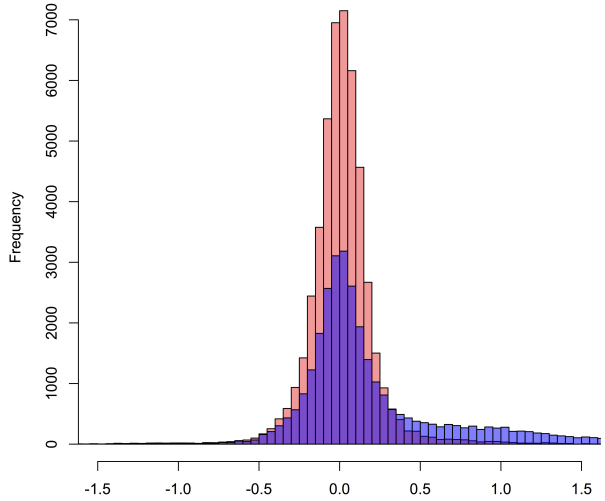

Figure 2: AVG error histogram of F2 estimated by ASSP (see Figure 1 for details).

hypothesis that the different voiced/unvoiced segmentations of the formant trackers would have a significant impact on evaluation results is therefore falsified.

6.3 Test Conditions VOWELS

Table 2 shows RMSE errors for three vowels [i:] (5694 frames), [u] (1373 frames) and [ʌ] (2671 frames).

		F1		F2		F3	
		f	**m**	**f**	**m**	**f**	**m**
SNACK	**[i:]**	91	**82**	372	**234**	239	**373**
	[u]	73	**54**	224	**163**	300	**421**
	[ʌ]	143	**105**	190	**155**	276	**364**
ASSP	**[i:]**	76	**66**	744	**245**	342	**200**
	[u]	65	**48**	280	**154**	317	**151**
	[ʌ]	117	**94**	347	**139**	634	**215**
PRAAT	**[i:]**	**62**	254	**216**	192	**190**	276
	[u]	**57**	469	**127**	300	**225**	425
	[ʌ]	**91**	145	**99**	207	**205**	281
DEEP	**[i:]**	97	80	268	185	226	167
	[u]	92	66	169	124	348	124
	[ʌ]	132	109	144	130	264	194

Table 3: RMSE errors in vowel segments [i:], [u] and [ʌ]

RMSE results are slightly better for vowels than across all phoneme classes which is not surprising. Again, parameter setting speaker's sex not matching speech input cause large error margins (up to 744Hz RMSE error for female speakers in ASSP). The advantage of DEEP against LPC-based trackers noted in the NORM test condition is not visible here. The low centralized vowel [ʌ] seems to be more difficult to track than the high vowels [i:] and [u]; one possible explanation is that reduced/centralized vowels in American English are often labelled with [ʌ] and are therefore more often hypo-articulated than [i:] and [u]. Hypo-articulation correlates with lower formant amplitudes and higher bandwidths which in turn makes the tracking of the formant frequency more difficult.

7. Conclusion

Four open source formant trackers, three LPC-based and one based on Deep Learning, were evaluated on an American English data set. The three traditional LPC-based formant trackers performed similarly well, when their respective parameter sets matched the sex of the tracked speaker; otherwise results were sometimes heavily skewed in one direction which could easily lead to the misinterpretation of tracker results. The average performance in terms of RMSE (210Hz) was about two times higher than reported comparable inter-labeller agreement error for human labellers on the same data set (96Hz, Deng et al 2006). The SNACK formant tracker turned out to be robust against wrong speaker sex settings, but sometimes produced default values as output (F1=500Hz, F2=1500Hz, ...) without warning; these could be theoretically misinterpreted as formant measurements. The Deep Learning tracker appears to out-perform traditional LPC-based methods in general but not when tested on vowels only. However, since the Deep Learning tracker was trained on the training set of the same speech corpus used for testing, this finding will probably not hold for other data sets (and especially not for other languages).

Some take home messages when dealing with formant trackers:

- LPC-based formant trackers show about 2 times less precision than human labellers.
- if possible, adjust your tracker to the sex of the target speaker.
- check for and remove default output values (repetitions of exactly the same value).
- if a histogram of tracker results shows more than one peak, this could be an indication of formant confusion (e.g. F4 is sometimes recognized as F3); a possible solution is to increase the number of formants or reduce the spectral range (depending on the tracker algorithm)
- expect less reliable results in centralized vowels (and in hypo-articulated speech in general).

8. Bibliographical References

Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.23, retrieved 2017-04-17 from http://www.praat.org/

Childers, D.G. (1978). Modern spectrum analysis. IEEE Press Selected Reprint Series, Hoboken, NJ: John Wiley & Sons Inc.

Dissen (2016). Formant Estimation and Tracking using Deep Learning. In: Proceedings of the INTERSPEECH, pp. 958 – 962.

Garofolo, J.S. & Lamel, L. & Fisher, M.W. & Fiscus, J. & S. Pallett, D.S. & Dahlgren, N.L. & Zue, V. (1992). TIMIT Acoustic-phonetic Continuous Speech Corpus. Philadelphia, PA: Linguistic Data Consortium.

Kammoun, M.A. & Gargouri, D. & Frikha, M. & Hamida, A.B. (2006). Cepstrum vs. LPC: A Comparative Study

for Speech Formant Frequencies Estimation. In: GESTS Int'l Trans. Communication and Signal Processing,Laboratoire d'Electronique et de la Technologie de l'Information (LETI), Vol. 9, pp. 87-102.

Markel, J.E. & Gray, A.H. (1982). Linear Prediction of Speech. New York, NY: Springer .

Rabiner, L.R. & Schafer (1970). System for automatic formant analysis of voiced speech. JASA Vol 47, pp. 634-648.

Rose, Ph. (2003). Forensic Speaker Identification. International Forensic Science and Investigation, CRC Press.

Kåre Sjölander (2017). Snack-Sound-Toolkit, retrieved 2017-04-17 from http://www.speech.kth.se/snack.

Keshet, J. (2017). DeepFormant, retrieved 2017-04-30 from https://github.com/MLSpeech.

Scheffer, M. (2017). Advanced Speech Signal Processor (libassp), retrieved 2017-04-17 from http://www.sourceforge.net/projects/libassp.

Winkelmann, R. & Harrington, J. & Jänsch, K. (2017). EMU-SMDS: Advanced speech database management and analysis in R. Computer, Speech & Language, 45 (2017), pp. 392-410.

9. Language Resource References

Deng, L. & Cui, X. & Pruvenok, R. & Huang, J. & Momen, S. & Chen, Y.N. & Alwan, A. (2006). A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. In: Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing.