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 showed that all three LPC-based trackers had comparable RSME error results that are about 2 times the inter-labeller error of human labellers. Tracker results were biased considerably (in average too high or low), when the parameter settings of the tracker were not adjusted to the speaker's sex. Deep Learning appeared 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'). Currently, a small number of open source formant trackers is widely used by speech scientists and speech engineers. A number of earlier studies evaluated selections of these formant trackers on selected vowels, on words spoken in isolation and/or on speech of a small number of speakers: Derdemezis et al. 2016 evaluated 4 LPC-based formant trackers on 4 different vowels in isolated words uttered by 8 speaker groups; Deng et al, 2006 evaluated their own formant tracker compared to WaveSurfer based on VTR-TIMIT; Harrison, 2004 compared three LPC-based algorithms on data of two speakers. But to our knowledge there exists no objective evaluation of the quality of more than two algorithms based on the same data set comprising a reasonable number of speakers of both genders and fluent speech. Partly this is due to the fact that a manually controlled reference corpus of fluent speech 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, at least for US American English.

Aside from the direct comparison of different formant trackers as presented in this study there exist numerous studies that compare the out-come of a single formant tracker to one or more expert annotations, or evaluate the influence of methodological parameters such as the time point of measurement (Kendall & Vaughn 2015), the number of formants or LPC order (Vallabha & Tuller, 2002; Harrison 2004), the signal quality (Rathcke et al. 2016), or even the outcome of different expert groups using the same formant tracker but different methodologies (Duckworth et al. 2011). Derdemezis et al (2016) investigated the influence of several measurement parameters; they also give a very detailed discussion of existing studies regarding parameter manipulation (see also Burris et al, 2014). Most of these earlier studies are in agreement that the quality of formant trackers' results can be improved by adjusting the parameters of the algorithm to the given task, i.e. dependent on the algorithm itself, the age, gender and health of speakers, the point of measurement, and the quality of the recording. On the other hand, many authors in the literature agree that, even when given methodological rules that may improve format tracker output, in practical terms most researcher tend to use formant trackers with their respective default settings and do not adjust tracking parameters as advised.

In this study we therefore do not give any methodological recommendations for the four tested formant trackers but rather compared the out-come when using default tracker parameters, and the impact caused by the voiced/unvoiced detection, the speaker genders and three vowel classes of 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 the formant estimation algorithm is 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 (= the formant tracker) is influenced by the performance of both algorithms.

In this study four formant trackers have been 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). Contrary to the three other formant trackers this algorithm produces formant frequency estimates at all time points, i.e. there is no voiced/unvoiced detection.

PRAAT, SNACK and ASSP 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 (Deng et al, 2004). The subcorpus contains 282 phonetically compact (SX in TIMIT terminology) and 234 phonetically rich sentences (SI), but no dialectal speech (SA). The speech was first analysed by the formant tracker 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 were 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 formant tracker DEEP was trained on parts of VTR-TIMIT, the following evaluations of DEEP were only performed on the test part of VTR-TIMIT (a subset with 8f and 16m speakers). We did not restrict the tests of the LPC-based trackers on this subset (this would have compromised the statistical power of the analysis considerably), since results are not comparable between DEEP and the remaining algorithms anyway: DEEP has been trained on the training part of VTR-TIMIT and is therefore language- and corpus-dependent, while the three other formant tracker algorithms are language- and corpus-independent.

5. Evaluation Methodology

Quality Measures

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

- **RSME**: root mean squared error calculated over the complete recording to quantify overall performance (zero being perfect match between formant tracker output and reference).

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

- **AVG**: average difference between reference and formant tracker output calculated over the complete recording to indicate systematic errors: a positive value indicates that the tracker tends to calculate formant estimates in average too low; a negative value indicates formant estimates are too high; zero indicates a perfect symmetry of errors around the annotation reference. Please note that the AVG *value* does not give any information about the quality.

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 formant tracker DEEP 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 and chosen parameter settings (defaults are marked with an 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 conditions

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 three 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 VTR-TIMIT recordings when tracker output is to be evaluated. If a tracker did not deliver `real' results within these defined areas, these `fake' or zero results were excluded. For instance, SNACK outputs 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**: trackers decided individually for which portions of the recording formant values were produced. Again, detectable `fake' values were 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. Since the formant tracker DEEP does not perform a voiced/unvoiced detection and PRAAT uses the same voiced/unvoiced detection as in the NORM test condition, only SNACK and ASSP results were evaluated in this test condition.
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. The three selected American vowels can be seen as representative sounds for high front, high back and low vowels.

6. Results

In the following RMSE and AVG results for all test conditions (see section 5) are presented.

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 increased with formant order; the relative errors were in about the same range for all three formants. Errors of the DEEP tracker were 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. The formant tracker PRAAT showed systematic errors with male speakers, while SNACK and ASSP performed better on male than on female speakers. However, when 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 was in the same range (underlined values in Table 2). It is unlikely that these differences between formant trackers are significant, considering that averaged human inter-labeller errors on the reference were in the range of 96Hz (see section 4). But it is clear that the LPC-based trackers in average (210Hz error) performed significantly worse than human labellers (about 2 times worse).

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

Table 2: RMSE errors in the test condition NORM (underlined values: formant tracker parameters match speaker sex).

Looking at averaged AVG measures in Table 3, the main result of the evaluation was that ASSP tended to underestimate formant frequencies F2/F3 for female speakers (200Hz too low), while PRAAT did the opposite, i.e. it overestimated all three formant frequencies for male speakers (120Hz too high). The averaged AVG errors for SNACK and DEEP were quite balanced especially for lower formants F1/F2.

	F1		F2		F3	
	f	m	f	m	f	m
SNACK	23	**<u>-9</u>**	1	**<u>-8</u>**	-45	**<u>-102</u>**
ASSP	-12	**<u>-14</u>**	187	**<u>11</u>**	216	**<u>24</u>**
PRAAT	**<u>-38</u>**	-116	**<u>11</u>**	-114	**<u>-13</u>**	-188
DEEP	86	56	5	-68	116	-26

Table 3: AVG errors in the test condition NORM rounded to two decimal places (underlined values: formant tracker parameters match speaker sex).

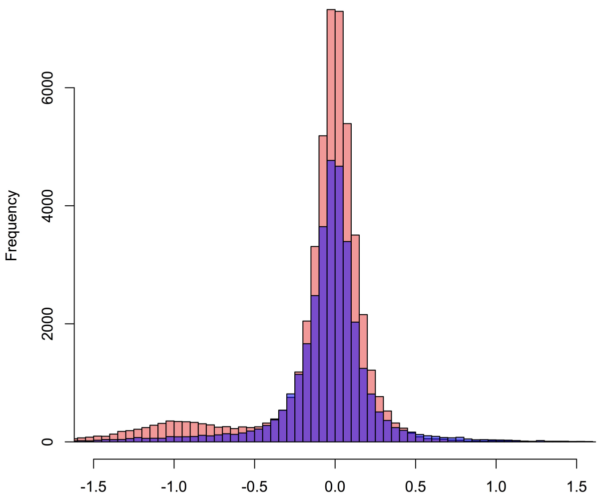

Figure 1: AVG error histogram of F3 (in kHz) estimated by SNACK; positive AVG error values denote estimates lower than reference, and vice versa; female speakers are light blue, male speakers are pink, overlap dark blue.

Selected AVG Histograms

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

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

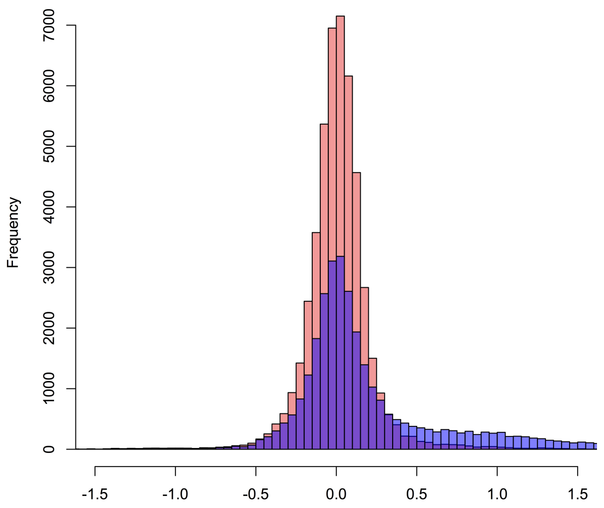

Figure 2: AVG error histogram of F2 (in kHz) estimated by ASSP (see Figure 1)

Fig. 3 displays the histogram of AVG errors for the first formant by the tracker PRAAT. One can clearly see that measurements for male Speakers (pink) were consistently overestimated (AVG error negatively skewed), and that formant tracker parameters were optimized for female speakers (light blue) which yielded AVG errors around zero.

6.2 Test Condition DEFAULT

The RMSE and AVG measures obtained in test condition DEFAULT for the formant trackers SNACK and ASSP were almost congruent with the results in the test condition NORM. Bonferroni corrected t-tests on *alpha = 0.01* applied on speaker aggregated errors RMSE and AVG (to avoid repeated measures) showed no significant changes in both quality measures and for the three first formants between test condition NORM and DEFAULT, except for the RMSE measure of the first formant F1 in formant tracker ASSP (but contrary to our expectation the error *increased* for the DEFAULT condition in this case). In total, when evaluating SNACK and ASSP the normalisation of pitch detection seemed not to be a significant factor (for brevity we do not show the error measure tables here). The hypothesis that the different voiced/unvoiced segmentations of the formant trackers would have a significant impact on evaluation results has therefore been falsified.

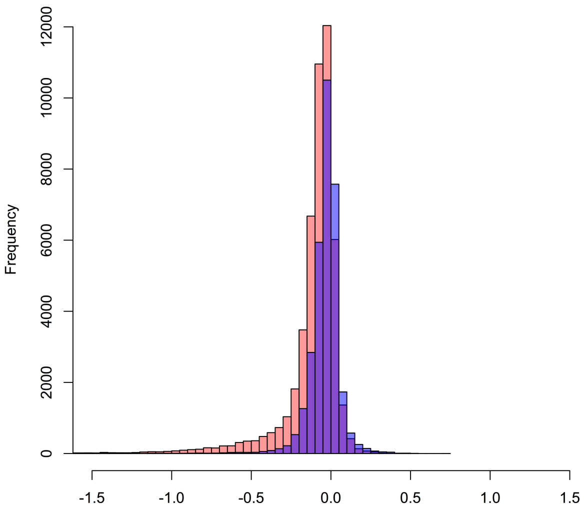

Figure 3: AVG error histogram of F1 (in kHz) estimated by PRAAT (see Figure 1 for details)

6.3 Test Conditions VOWELS

Table 4 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 4: RMSE errors in vowel segments [i:], [u] and [ʌ] (underlined values: formant tracker parameters match speaker sex).

RMSE results were significantly better for vowel segments than across the total recording which is not surprising. Again, formant tracker parameters that do not match speaker's sex caused larger errors (e.g. 744Hz RMSE error on F2 for female speakers in ASSP). The advantage of DEEP against the LPC-based trackers noted in the NORM test condition was not as prominent here. The low centralized vowel [ʌ] seemed 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. Another explanation is that F1 and F2 tend to be close together for [ʌ] and might therefore be harder to separate by the formant tracker.

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; if not filtered these could be misinterpreted as formant measurements. The Deep Learning formant tracker appeared to out-perform traditional LPC-based methods in general but not when tested on vowels only. However, since the Deep Learning formant 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.
- whenever 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) and in lower vowels.

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/

Burris, C. & Vorperian, H.K. & Fourakis, M. & Kent, R.D. & Bolt, D.M. (2014). Quantitative and descriptive comparison of four acoustic analysis systems: vowel measurements. Journal of Speech, Language and Hearing Research, 2014 Feb;57(1):26-45. doi: 10.1044/1092-4388(2013/12-0103).

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

Derdemezis, E. & Vorperian, H.K. & Kent, R.D. & Fourakis, M. & Reinicke, E.L. & Bolt, D. M. (2016). Optimizing Vowel Formant Measurements in Four Acoustic Analysis Systems for Diverse Speaker Groups. American Journal on Speech & Language Pathology, 2016 Aug; 25(3): 335–354. doi: 10.1044/2015_AJSLP-15-0020.

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

Duckworth, M. & McDougall, K. & de Jong, G. & Shockey, L. (2011). Improving the consistency of formant measurement. International Journal of Speech, Language & Law 18, 35-51.

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.

Harrison, P. (2004). Variability of formant measurements. MA Dissertation. York, UK:University of York.

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.

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

Kendall, T. & Vaughn, C. (2015). Measurement variability in vowel formant estimation: A simulation experiment. In: Proc. of the International Conference on Phonetic Sciences 2015, Glasgow.

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

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.

Rathcke, T. & Stuart-Smith, J. & Torsney, B. & Harrington, J. (2016). The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetic real-time studies. Speech Communication 86 (2016), pp. 24-41.

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

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

Vallabha, G., Tuller, B. (2002). Systematic errors in the formant analysis of steady-state vowels. Speech Communication 38, 141-160.

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.