INTERSPEECH 2009 BRIGHTON

An Evaluation of Formant Tracking methods on an Arabic Database

Imen Jemaa [1, 2], *Oussama Rekhis* [1], *Kaïs Ouni* [1] *and Yves Laprie* [2]

[1] Unité de Recherche Traitement du Signal, Traitement de l'Image et Reconnaissance de Formes (99/UR/1119)
Ecole Nationale d'Ingénieurs de Tunis, BP.37, Le Belvédère 1002, Tunis, Tunisia
imen_jemaa@yahoo.fr,oussamarekhis@gmail.com and kais.ouni@enit.rnu.tn
[2]Equipe Parole, LORIA-CNRS – BP 239 – 54506 Vandœuvre-lès-Nancy, France
Yves.Laprie@loria.fr

Abstract

In this paper we present a formant database of Arabic used to evaluate our new automatic formant tracking algorithm based on Fourier ridges detection. In this method we have introduced a continuity constraint based on the computation of centres of gravity for a set of formant candidates. This leads to connect a frame of speech to its neighbours and thus improves the robustness of tracking. The formant trajectories obtained by the algorithm proposed are compared to those of the hand edited formant database and those given by Praat with LPC data.

Index Terms: Arabic database, phonetic annotation, formant labeling and formant tracking

1. Introduction

It is well known that the spectral maxima of voiced speech, i.e. formants, play an important role in the identification of speech sounds. Indeed, robust formant tracking is utilized to identify vowels [1] and other vocalic sounds [2], to pilot formant synthesizers and, in some cases, to provide speech recognition with additional data. Although automatic formant tracking has a wide range of applications, it is still an open problem in speech analysis. Especially, when anti-formants of consonants are present, resonances frequencies, i.e. formants, are often hidden.

Due to the importance of the vocal tract resonances, numerous works have been dedicated to develop automatic formant tracking methods for estimating formant frequencies from the acoustic signal. Most of these methods are based firstly on the detection of the LPC roots [3] as initial estimates of formant frequencies. Results of many of these methods have been used in speech processing applications. However, there has been a conspicuous lack of databases that are needed for quantitative evaluation of automatic formant tracking, especially for the Arabic language. The construction of a corpus annotated in terms of formants is not an easy task. Indeed formants are often not visible in consonant segments and semi-vowels as well. Additionally, formants are sometimes very close from each other. The database was revised by human experts once the formant annotations were completed.

In this paper we describe our new automatic formant tracking algorithm based on the detection of Fourier ridges which are the maxima of spectrogram. This algorithm uses a continuity constraint by calculating the centre of gravity for a set of frequency formant candidates. Then, to evaluate the proposed algorithm we compare it to other automatic formant tracking methods using the labelled database as a reference.

This paper is presented as follows. In section 2 we present the Arabic database, in section 3, the different stages of the manual formant annotation process, in section 4, the description of the proposed automatic formant tracking algorithm, in section 5 the results obtained by comparing the proposed Fourier ridges algorithm with other automatic formant tracking methods. Finally, we give some perspectives in section 6.

2. The Arabic database

Given the interest of studies dedicated to formant tracking applied to the Arabic language and the lack of public formant Arabic database, we have decided to record and label in terms of formants a corpus of Standard Arabic pronounced by Tunisian speakers which will be publically available.

To build our corpus, we used a list of phonetically balanced Arabic sentences proposed by Boodraa and al. [4]. This corpus is also based on studies of Moussa [5] about the standard Arabic language to respect the frequency occurrence of each phoneme. This corpus follows a number of syntax, grammar and modal rules encountered while reading an Arabic text. Hence, it covers the whole phonetic and phonological behavior of the standard Arabic language. Most sentences of this database are extracted from Koran and Hadith. It consists of 20 lists of 10 short sentences. Each list of the corpus consists of 104 CV (C: consonant and V: vowel), i.e. 208 phonemes [4].

We recorded this corpus in a soundproof room. It comprises ten Tunisian speakers (five male speakers and five female speakers). The signal is digitized at a frequency of 16 kHz. This corpus contains 2000 sentences (200 different sentences pronounced by every speaker) either affirmative or interrogative. In this way, the database presents a well balanced selection of speakers, genders and phonemes. All the utterances of the corpus contain rich phonetic contexts and thus are a good collection of acoustic-phonetic phenomena that exhibit interesting formant trajectories [6]. In order to get accurate formant annotations, we have to verify every frequency value of formants with respect to phonemes uttered. Thus, to prepare our database, we phonetically annotated all the sentences of the corpus by hand using Winsnoori software [7]. A screenshot of this tool is shown in Fig.1. We used the code SAMPA [8] for this phonetic annotation task. In this stage of the work, we met a lot of difficulties to fix phonemes boundaries especially for consonants. Once the annotation phase has been completed, the corpus has been reviewed by three phonetician experts to correct certain mistakes.

3. Formant track labeling

To facilitate the process of formant trajectory labeling in the database preparation, we first obtained a set of frequency formant candidates provided by the LPC roots [3] using Winsnoori [7]. Based on these initial estimative values, we drawn formant trajectories by hand by selecting relevant trajectories. Fig.1 shows an example of a sentence pronounced by a female speaker to illustrate the formant labeling process and results. We labeled the first three formants (F1, F2 and F3) every 4 ms and recorded them for each sentence of the database. The LPC order used is 18 and the temporal size of the window analysis is 4 ms to have a wideband spectrogram which shows the evolution of the formant trajectories better. More difficult situations arise when there are too many LPC candidates which are close to each other for two formant trajectories. But most of the difficulties arise for the frames where there is a lack of spectral prominences or when spectral prominences do not coincide with predicted resonances for consonantal segments. In this case, nominal consonant specific values are provided [9][10].

Finally, in order to verify the accuracy of the formant trajectories for each sentence, we have synthesized the sound with the corresponding three formant frequencies using Klatt synthesizer implemented in Winsnoori [7] to check whether the synthesized sentence matches the original one well. The evaluation was subjective though, since the authors were the only judges of the result quality.

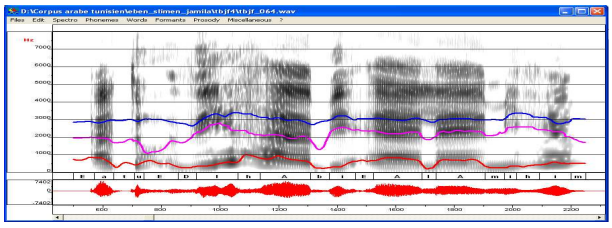

Figure 1: *Formant trajectories labeled (F1, F2 and F3) by hand and manually phonetic annotation for the sentence" أَتُؤْذِيهَا بِآلامِهِمْ؟ "3atu3Di:ha: bi3a:la:mihim" (Would you harm her feelings with their pains) pronounced by a female speaker.*

4. An automatic formant tracking method using Fourier ridges

The block diagram presented in Fig.2 describes the main steps of the proposed algorithm. Each element of the block diagram is briefly described below.

4.1. Preprocessing

The sampling frequency of the database is Fs=16 kHz. Since we are interested in the first three formants, we re-sampled the speech signal at Fs=8 kHz in order not to take into account formant candidates above 4 kHz and to optimize the computation time. Then, to accentuate high frequencies, a first order pre-emphasis filter of the form $1-0.98z^{-1}$ is applied on the speech signal.

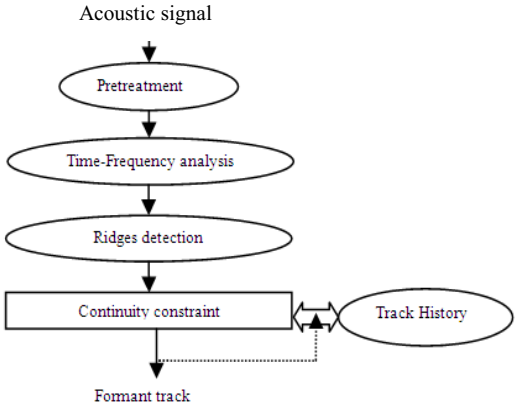

Figure 2: *Block diagram of the proposed formant tracking algorithm.*

4.2. Time-Frequency analysis

The time-frequency analysis of the signal is carried by the module to the square of the windowed Fourier transform to obtain the spectrogram of the signal. The spectrogram used here is a wideband spectrogram witch smoothes the spectral envelope of the signal to show the temporal evolution of formants.

4.3. Detection of ridges

Since formant frequencies vary as a function of time, they were assimilated to the instantaneous frequencies of the signal. Thus, the algorithm calculated all instantaneous frequencies of the input signal which were considered as local maxima of the spectrogram. In the following, we show how the instantaneous frequency was validated like local maximum of the spectrogram (for detailed proof, see [11]). We generated the family of time-frequency atoms, .i.e windowed function, of the windowed Fourier transform noted $g_{u,\xi}$ by time translations and frequency modulations of a real and symmetric window *g(t)*. This atom has a centre frequency ξ and is symmetric with respect to u (translation factor). The windowed Fourier transform $Sf(u,\xi)$ is carried by the correlation of this atom with the input signal *f*.

It was shown in [11] that the instantaneous frequency $\phi'(t)$ of f which is the positive derived phase of the signal, is related to the windowed Fourier transform $Sf(u,\xi)$ if $\xi \geq 0$ (Eq.1).

$$Sf(u,\xi) = \frac{\sqrt{s}}{2} a(u) \exp^{i(\phi(u)-\xi(u))} \times \left[\hat{g}\left(s\left[\xi - \phi'(u)\right]\right) + \varepsilon(u,\xi) \right] \quad (1)$$

Where s is a scaling which has been applied to the window Fourier g, $\hat{g}$ the Fourier Transform, FT, of g, $a(t)$ the analytic amplitude of f and $\varepsilon(u,\xi)$ a corrective term.

Since $|\hat{g}(\omega)|$ is maximum at $\omega = 0$, Equation (1) shows that for each u, the spectrogram $|Sf(u,\xi)|^2$ is maximum at its centre frequency $\xi(u) = \phi'(u)$. As a result, the instantaneous frequency is well validated as a local maximum of the spectrogram. The windowed Fourier ridges are the maxima of the spectrogram at the FT points $(u, \xi(u))$. In each temporal window analysis, the algorithm thus detects all local maxima

of the FT representation assimilated at Fourier ridges in the $(u, \xi(u))$ plane. We have noticed that Hamming window gives better results in term of formant detection resolution than other windows [12]. We thus use a 4 ms window with an overlap equal to 50 %.

In this work tracking only concerns the first three formants, so in each temporal window analysis, we have split the set of frequency formant candidates detected previously into three wide bands each of them corresponding to a formant. Then, a parabolic interpolation is applied and ridges below a threshold are removed because there may be artefacts (for example, "shadows" of other frequencies produced by the side-lobes of the window of the Fourier transform analysis or instantaneous frequencies specific to the fundamental frequency F0 [12]). Then, we calculate frequencies corresponding to the remaining ridge points. These frequencies are formant candidates that might be chosen to form formant trajectories. At this point, a set of frequency candidates is available for each formant.

4.4. Continuity constraint

It is considered that in general, formants vary slowly as a function of time what leads to impose a continuity constraint in the process of selecting formant frequencies from the set of candidates. For the other algorithms based on LPC spectra, the continuity constraint used for each formant trajectory is the moving average of the LPC roots over its respective frequency band [7] [13]. In this algorithm we propose the calculation of the centre of gravity for a set of frequency formant candidates, detected by the ridge detection stage, as a continuity constraint between signal frames. Since the detection of ridges gives several candidates close together for one formant, the idea is to calculate the centre of gravity of the set of candidates located in the frequency band associated to the formant considered. The resulting frequency f is given by

$$\bar{f} = \frac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i} \qquad (2)$$

where f_i is the frequency of the i^{th} candidate and p_i its spectral energy.
Considering centres of gravity instead of isolated candidates allows smooth formant transitions to be recovered.

5. Results and Discussion

We used our formant database as a reference to verify the accuracy of the proposed algorithm and to compare it with other automatic formant tracking methods. This evaluation has been conducted on the corpus presented above for the Fourier ridge algorithm and the popular open source Praat system [13]. To enable a visual comparison, Figures 1, 3 and 4 show the automatic formant (F1, F2 and F3) tracking results for the same example sentence"أَتُؤْذِيهَا بِآلامِهِمْ؟"("3atu3Di:ha: bi3a:la:mihim" which means "Would you harm her feelings with their pains") pronounced by a female speaker. The first figure is the reference obtained by hand, and the other two are respectively the result obtained by our algorithm and that of the Praat system. The default LPC order used in Praat is 16. The size of the window analysis is 25 ms. It can be seen that for most of the vocalic portions in the utterances where the "dark/high energy" bands in the wideband spectrograms are clearly identifiable, both automatic trackers give accurate results. Exceptions are occasional errors in F2 and especially F3 in the long vowel /I/ for the two algorithms, contrary to the short vowel /i/ where both methods reach correct results approximately. These errors occur mostly during relatively rapid formant transitions.

To quantitatively evaluate the proposed algorithm, we compare it with the LPC method of Praat using our formant database as a reference. So, the evaluation of each method consists in calculating the averaging absolute difference (Eq.3) and the standard deviation normalized with respect to manual reference values (Eq.4) for every formant trajectory (F1, F2 and F3). We thus examined results obtained for the short vowel /a/, which is the most voiced one in the Arabic language [9] [10], within the syllable CV. Table 1 shows the results obtained on the vowel /a/ preceded by one consonant from every phonetic class for each formant trajectory (F1, F2 et F3) and for both tracking method. The CV occurrences were taken from the two following sentences pronounced by one male speaker: "عَرَفَ وَالِيًا وَقَائِدًا " ("εarafa wa:liyan wa qa:3idan" which means "He knew a governor and a commander") and "هِيَ هُنَا لَقَدْ آبَتْ"("hiya huna: laqad 3a:bat" which means "She is here and she was pious").

$$Diff = \frac{1}{N} \times \sum_{p=1}^{N} \left| F_r(p) - F_c(p) \right| \text{ Hz} \qquad (3)$$

Where F_r is the reference frequency, F_c the calculated frequency corresponding to both formant tracking methods and N the total number of points of each formant trajectory.

$$\sigma = \sqrt{\frac{1}{N} \sum_{p=1}^{N} \left(\frac{\left| F_r(p) - F_c(p) \right|}{F_r} \right)^2} \qquad (4)$$

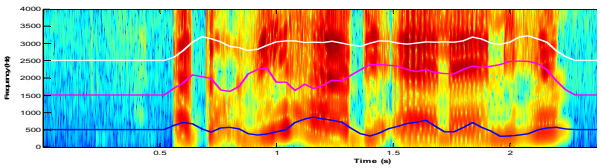

Figure 3: *Automatic formant trajectories (F1, F2 and F3) obtained by the Fourier ridge algorithm for the sentence "3atu3Di:ha: bi3a:la:mihim"(Would you harm her feelings with their pains).*

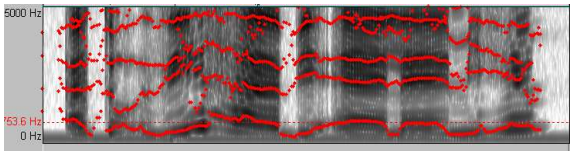

Figure 4: *Automatic Formant trajectories by LPC (Praat) for the sentence "3atu3Di:ha: bi3a:la:mihim" (Would you harm her feelings with their pains).*

The comparison of the errors shown in Table 1 shows that, globally, there is no big difference in terms of errors between the two automatic formant tracking methods except in some cases when the vowel /a/ is preceded 1) by a semi-vowel for F2 and F3, 2) by a tap for F3, 3) by a voiced plosive for F1 and F2. In these cases the algorithm proposed presents results close to the reference, and better results than the LPC method especially for F1 and F2 (except for F2 when the vowel /a/ is preceded by a voiceless plosive).

Table 1. *Formant tracking errors measured by averaging absolute difference (Hz) and standard*

deviation (%) between the reference and estimated values for the short vowel /a/ within the syllable CV with different types of consonant.

Formant Tracks		LPC			Fourier Ridges		
		F1	F2	F3	F1	F2	F3
Plosive voiced :	**Diff**	119	93	212	78	29	285
/d /	**σ**	42	37	65	18	6	71
Plosive voiceless:	**Diff**	71	58	67	75	110	17
/q/	**σ**	16	11	14	17	24	3
Fricative voiced	**Diff**	49	48	36	28	49	33
/H/	**σ**	8	7	6	4	8	6
Fricative voiceless:	**Diff**	44	31	79	16	40	36
/f/	**σ**	14	8	24	4	11	8
Nasal:	**Diff**	97	30	101	73	47	166
/m/	**σ**	22	7	25	15	13	46
Latéral:	**Diff**	76	34	45	35	76	105
/l/	**σ**	24	8	11	8	15	24
Tap:	**Diff**	57	68	246	34	78	166
/r/	**σ**	13	16	59	8	18	34
Semi-vowel:	**Diff**	98	91	140	92	32	89
/w/	**σ**	31	49	52	18	7	20

We also notice that in most cases, the Fourier ridge algorithm presents large difference errors to the reference for F3. Tables 2 and 3 present formant tracking errors measured by averaging absolute difference and standard deviation between the reference and estimated values for each type of vowel, i.e. short vowels (/a/,/i/ and /u/) and long vowels (/A/, /I/ and /U/) pronounced respectively by two male speakers and two female speakers. Tests have been performed on four sentences: "هِيَ هُنَا لَقَدْ أَبَتْ", "عَرَفَ وَالِيًا وَقَائِدٌ. أَنُوْذِيهَا بِآلامِهِمْ؟", cited above, and a last one "أَسَرُونَا بِمُنْعَطَف"("Ɜasaru:na: bimunɛatafin" which means "They captured us at a bend").

Table 2. *Formant tracking errors measured by averaging absolute difference (Hz) and standard deviation (%) between the reference and estimated values for each type of vowel pronounced by two different male speakers.*

		MS1						MS2					
		LPC			Fourier Ridges			LPC			Fourier Ridges		
		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
a	Diff	44	31	79	16	40	36	34	44	152	17	18	103
	σ	14	8	24	4	11	8	9	12	49	5	4	25
A	Diff	52	91	89	38	71	82	58	114	63	79	82	34
	σ	14	45	29	9	14	18	13	23	14	16	18	7
i	Diff	35	49	58	28	25	91	28	53	161	26	190	192
	σ	12	18	20	9	9	28	12	20	73	10	64	69
I	Diff	42	67	64	170	198	12	64	47	83	45	150	166
	σ	13	19	21	46	54	4	20	14	25	15	41	48
u	Diff	57	82	89	118	95	99	26	55	90	91	80	28
	σ	15	20	28	32	23	41	10	20	31	31	26	9
U	Diff	65	83	265	133	62	160	86	191	215	58	93	187
	σ	19	22	64	41	17	52	40	125	103	19	29	86

Table 3. *Formant tracking errors measured by averaging absolute difference (Hz) and standard deviation (%) between the reference and estimated values for each type of vowel pronounced by two different female speakers.*

		WS1						WS2					
		LPC			Fourier Ridges			LPC			Fourier Ridges		
		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
a	Diff	46	59	49	25	70	30	47	107	48	30	100	34
	σ	10	18	11	6	15	6	9	22	9	6	19	6
A	Diff	93	115	100	168	91	115	44	51	113	189	166	88
	σ	14	21	14	20	15	17	6	7	14	24	27	11
i	Diff	33	87	75	49	82	56	39	57	104	41	127	64
	σ	11	20	21	15	26	14	12	29	39	14	49	20
I	Diff	32	110	185	24	411	268	38	445	289	15	624	225
	σ	11	41	66	7	106	75	10	137	78	4	162	53
u	Diff	48	138	346	109	51	290	112	343	690	85	296	158
	σ	17	82	124	37	16	80	30	129	150	19	78	55
U	Diff	36	96	182	46	77	240	43	90	158	47	167	92
	σ	10	29	58	15	23	68	13	24	46	13	47	26

Table 2 shows that results are good especially for the vowels (/a/,/A/ et /i/) contrary to the (/I/,/u/ and /U/) probably because of their lower energy. We also notice that the proposed algorithm presents better results for the male speaker1 (MS1) than for (MS2). Table 3 shows that results are good for the vowels (/a/,/i/) for both tracking methods, especially for the female speaker 1 (WS1) contrary to (WS2) where there are some errors for F3 and F2 for both algorithms. However, results are not very good for the other vowels uttered by female speakers whatever the tracking method. Finally, it can be concluded that our algorithm presents a good reliability especially for sentences pronounced by male speakers.

6. Conclusion

In this paper, the development of a formant Arabic database is presented. Formant trajectories have been labeled by hand and checked by phonetician experts. This database is well balanced with respect to gender and phonetic contexts. Furthermore, we report in this paper an exploratory use of the database to quantitatively evaluate a new automatic formant tracking algorithm based on the detection of Fourier ridges. This algorithm provides accurate formant trajectories for F1 and F2 and results not as good for F3 in some cases. Our future work will target the improvement of this method especially for high frequency formants.

7. Acknowledgment

This work is supported by the CMCU (Comité Mixte Franco-Tunisien de Coopération Universitaire), Project 07G 1112.

8. References

[1] Thibault, F., "Formant Trajectory Detection using Hidden Markov Models", In Proc. Of Sound Processing and Control Lab, Montreal, Canada, 2003.

[2] Ali, J. A. M., Spiegel, J. V. D. and P. Mueller, "Robust Auditory-based Processing using the Average Localized Synchrony Detection", In Proc. of IEEE Trans. Speech and Audio Proc, 2002.

[3] McCandless, S., "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans, 22:135-141, 1974

[4] Boudraa, M., Boudraa, B. and Guerin, B., "Twenty Lists of Ten Arabic Sentences for Assessment", Act of Communication ACUSTICA, 86:870-882, 2000.

[5] Moussa, A. H., "Statical study of Arabic roots on Moɛjam Al-Sehah", Kuwait University, 1973.

[6] Deng, L.,"A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing", In Proc. of ICASSP, 2006.

[7] http://www.loria.fr/~laprie/WinSnoori/.

[8] http://www.phon.ucl.ac.uk/sampa/home.htm/.

[9] Ghazeli, S., "Back consonants and backing coarticulation in Arabic", PhD dissertation, University of Texas, Austin, 1977.

[10] Braham, A., "An Acoustic study of temporal organization in Arabic specific to Tunisian speakers", PhD dissertation,(written in Arabic), university of Manouba, Tunis, 1997.

[11] Mallat, S.,"A Wavelet Tour of Signal Processing", Academic Press, 1999.

[12] Châari, S., Ouni, K. and Ellouze, N.,"Wavelet Ridge Track Interpretation in Terms of Formants", In Proc. of INTERSPEECH-ICSLP, 1017-1020. Pittsburgh, Pennsylvania, USA, 2006.

[13] http://www.praat.org/