On the Way to Precise and Robust Formant Frequencies Tracking*

Boris Lobanov and Andrei Davydau

United Institute of Informatics Problems Nat. Ac. Of Sc. Of Belarus
Surganov Str. 6, 220012 Minsk, Belarus
lobanov@newman.bas-net.by

Abstract

The paper presents a short review of existing methods of formant analysis and an extensive bibliography on the given issue. A brief description of a new formant analysis method based on LSF-representation of the LPC spectrum is given. Preliminary results of the estimation of the suggested method efficiency are presented.

1. Introduction

Formant frequencies (or formants) are the best known and the most commonly used parameters in characterizing speech sounds. Formants represent the resonances in the human vocal tract during speech production and the dynamic structure of speech articulation in the domain of vocal tract resonance (VTR). Resonance frequencies of the human vocal tract are of fundamental importance in speech production and perception [1 - 6].

The temporal trends of CV and VC formant transitions (a classic concept known as formant "loci" [7] having perceptual relevance) help to capture the dynamic structure of long-range speech coarticulation and reduction [8-11]

Formants are associated with peaks, maximums or prominence in the smoothed power spectrum of the acoustic signal of speech. But in real speech not any spectral maximum is a formant, and, conversely, the formant is not always manifested as the spectral maximum. According to this acoustic definition, formants would "disappear" during complete consonantal closure, and may "split" or "merge" under other conditions when the peaks in the acoustic spectrum become ambiguous. In spite of the fact that the exclusive importance of the formants has been acknowledged by researchers already since more than half a century ago, the problem of formant analysis still remains largely unsolved, particularly, because of the existence of this kind of ambiguities. However, recently as a result of general progress in the development of methodology of speech signals analysis and recognition a great number of works have appeared which are devoted to formant analysis. Different methods of the speech signal parameterization – various modifications of FFT and LPC spectrum and cepstrum – as well as different methods of interpretation of the results of primary analysis – dynamic programming, hidden Markov models, mixtures of Gaussian, Kalman filtering and others – are currently used for solving the tasks of formant analysis (see: [12-40]). Methods of the formant analysis developed to the present time are successfully used in modern systems of speech recognition [41-43] and speaker verification [44-46].

However, there remains an unresolved problem of how to estimate the efficiency of this or that method. New methods of efficiency estimation suggested nowadays (since 2007) by the majority of authors make use of a recently released public database [47] of formant trajectories. The formant trajectories (or VTR) database comprises hand-corrected vocal tract resonance information for a subset of the TIMIT database. TIMIT is a large vocabulary database containing speech from eight US dialects. Although the VTR database provides the frequencies and bandwidths of the first four formants for 512 utterances, only the frequencies of the lowest three formants are hand-corrected. The database is split into testing (192 utterances from 24 male and female speakers) and training (324 utterances from 173 different speakers).

As a baseline for comparison of a newly suggested method a free Snack Toolkit [48] is often used. The tool estimates formant frequencies by solving for the roots of a 12_{th} order linear predictor polynomial. Dynamic programming is used to find the optimal formant tracks. For a given frame, all mappings of the complex roots to the estimated formant frequencies for the previous frame are calculated and a cost value, based on formant frequencies and bandwidths, is obtained. The optimum formant track is given by the path with the lowest value.

The organization of this paper is as follows. In Section 2, a short description of a new formant analysis method based on LSF-representation of a LPC spectrum is given. In Section 3, we present preliminary results of efficiency estimation by means of the suggested method and some discussion of its further improvements.

2. Method

Here we suggest a new method of formant analysis based on line spectrum frequencies (LSF) or pairs (LSP) representation of LPC. As one can see from the list of references [12-40], there has been no example of using the LSF representation of LPC for formant analysis. The concept of LSF was introduced by Itakura, but it remained almost dormant until its usefulness was re-examined in the latest speech coding standards. LSFs encode speech spectral information in the frequency domain and have been found to be capable of improving the coding efficiency more than other transformation techniques, especially when incorporated into predictive quantization schemes [49].

As is well known, the general spectrum of a speech signal caused by excitation of the vocal tract and radiation, is described by means of a linear system with the transfer function expressed by the following formula:

$$H(z)=\frac{1}{A(z)}=\frac{1}{1-\sum_{k=1}^{p}a_k z^{-k}} \qquad (1)$$

Where $\{a_k\}$ is a set of coefficients which are called LPC-parameters, and p is prediction order.

To receive *LSF*-coefficients, p zeros of function $A(z)$ are reflected on a united circle by means of two z-transformations $P(z)$ and $Q(z)$ of $(p+1)^{th}$ order:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \quad (2)$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \quad (3)$$

It follows from this, that

$$A(z) = \frac{P(z) + Q(z)}{2} \quad (4)$$

LSF- coefficients represent angular positions of roots $P(z)$ and $Q(z)$ on a united circle in a range $0 \le \omega_i \le \pi$.

They have the following properties:

– All roots $P(z)$ and $Q(z)$ lie on an unite circle;

– Roots alternate on an united circle, i.e. the following inequality is carried out:

$$0 < \omega_{p,1} < \omega_{q,1} < \omega_{p,2} < \omega_{q,2} < \ldots < \pi \quad (5)$$

LSF representation of LPC spectrum possesses the following important properties:

1) The distance between LSF defines the amplitude of spectral density;

2) The block from two or three close located LSFs shows the maximum presence in a spectrum, while located with large interval between LSF corresponding minima;

3) Generally, spectral sensitivity of each LSF is localized, i.e. at little change of one of LSFs the spectrum will change only in the vicinities of these LSF-parameters.

The first two of the above properties are taken as the basis in the development of our method of formant frequencies tracking. The third property is very important for the solution of speech coding problems.

In figure 1 different representations of the speech signal for the phrase *"Even I occasionally get the Monday blue!"* spoken by a female speaker are shown.

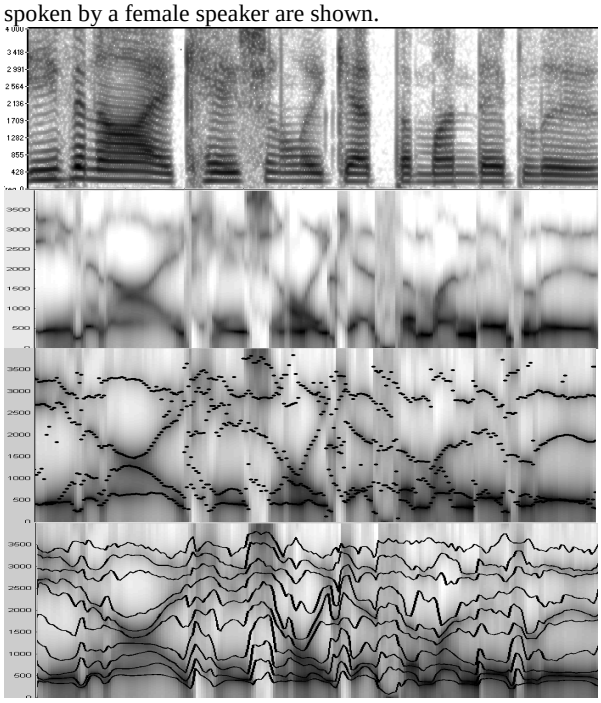

Figure 1 – Different representations of a speech signal. See from the top to the bottom: FFT spectrum (a), LPC spectrum (b), LPC spectrum+pole (c), LPC spectrum+LSF (d).

Each spectrogram was obtained for the signal from the TIMIT DB with a sampling rate of 8 kHz and with a window of 32 mc. Comparing 1a and 1b one can see that LPS-spectrum shows a more accurate picture of the formant movements than the FFT-spectrum. The poles of transfer function shown on LPC spectrum (1c) give additional information on the position of formants, though it is not always quite correct. It is apparent from 1d, that continuous LSF trajectories closely correlate with formant movements that can serve as additional information for formant frequencies tracking. In the method suggested we combine each of the three advantages of LPC speech signal representation, i.e. LPC spectrum, poles and LSF, in order to improve precision and robustness of formant frequencies tracking.

3. Software model

The general structure of the software model realized by the suggested method of formant analysis is shown in figure 2. It consists of four main blocks:

- LPC analysis block that calculates the spectrum {H(w)}, poles $\{p_i\}$ and line spectrum frequencies $\{LSF_i\}$ of the speech signal (see figure 3 for more details);
- Formant range detection block that determines the ranges $\{F\_rng_i\}$ of formants frequencies changes based on LSF information (see figure 4 for more details);
- Formants finding block that determine the formant frequencies F1, F2, F3 inside their ranges based on the spectrum {H(w)}, poles $\{p_i\}$ and line spectrum frequencies $\{LSF_i\}$ information (see figure 5 for more details);
- Tracks interpolation block that provides smooth connection of formant frequencies tracks on the unvoiced regions (see figure 6 for more details).

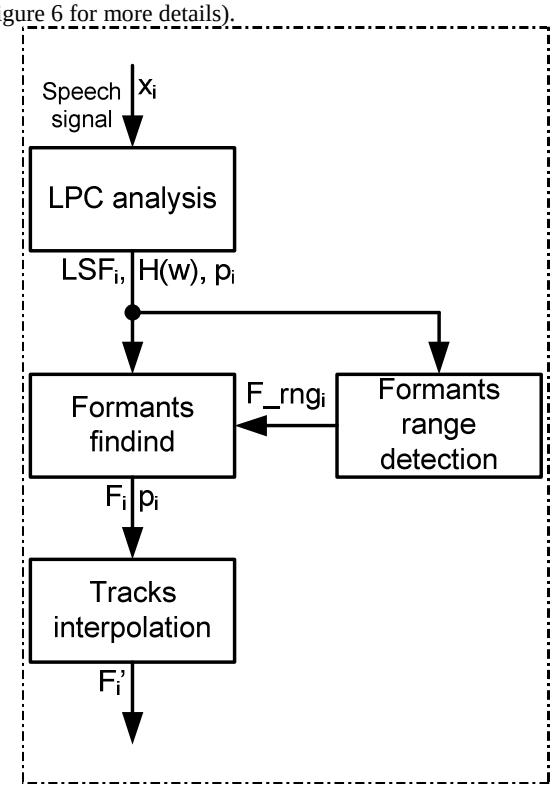

Figure 2 –General structure of formant analysis model

The software model is designed as a toolkit for investigation and includes a big number of adjustment parameters.

LPC analysis block (figure 3) provides at its output not only direct values of poles position $\{p_i\}$ and line spectrum frequencies $\{LSF_i\}$ but also their weighted values estimamted according to spectral density: wp_i and $wLSF_i$.

Adaptive formant range detection block (figure 4) based on a choice of an optimum set of the LSF_i which are the best borders of *F1, F2, F3* formant frequencies regions. It utilizes different formants range detection methods as is shown in figure 4.

Formants finding block provides calculation of three formant frequencies by using different types of LPC parameters or their combinations as is shown in figure 5.

Formant tracks interpolation block (figure 6) provides linear or cubic spline interpolation of formant frequencies tracks on the unvoiced regions by using pole weighted mean or maximum.

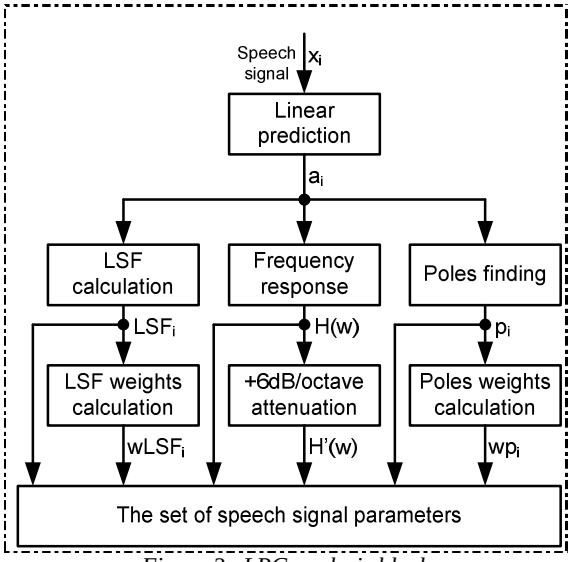

Figure 3 –LPC analysis block

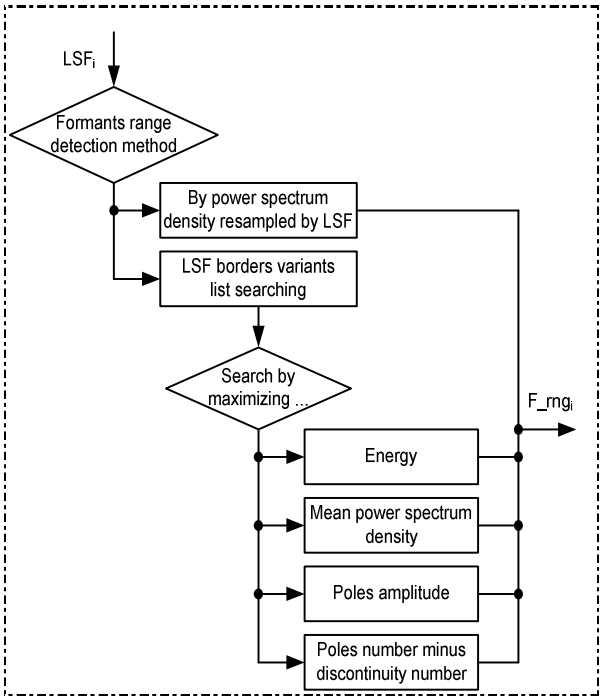

Figure 4 – Adaptive formant range detection block

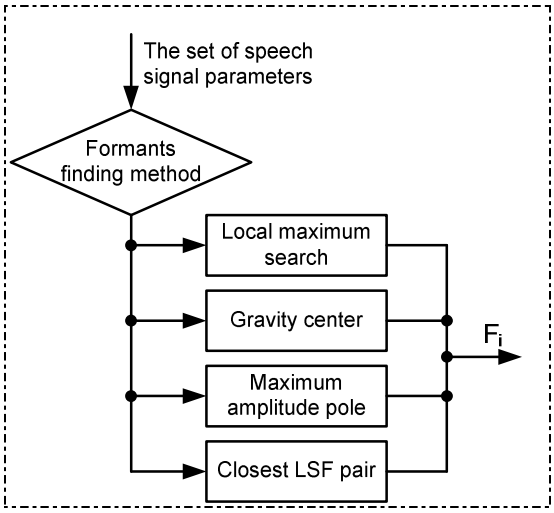

Figure 5 – Formants finding block

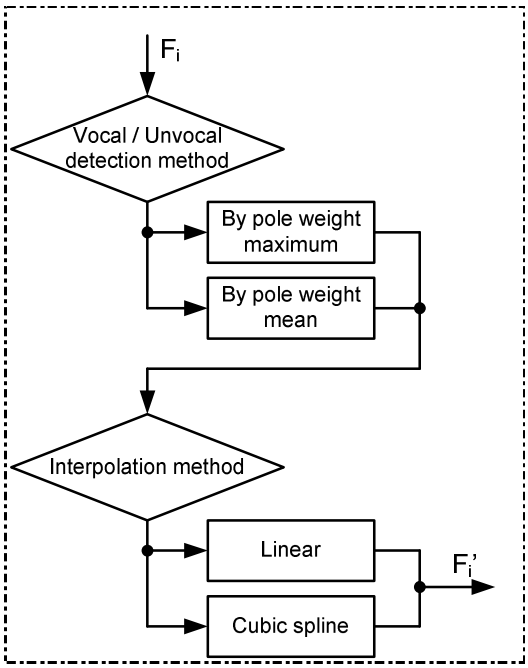

Figure 6 – Tracks interpolation block

4. Results and discussions

Here we will present the preliminary testing results of our model demonstrated by phrases taken from TIMIT DB spoken by different speakers [47]. The formant frequencies tracking result for the phrase "*A roll of wire lay near the wall*" spoken by three different male speakers is shown in figure 7. In figures 8 and 9 the formants trajectories for the same phrases and speakers as in figures 1 and 2 from paper [38] are shown for comparison.

In all figures the formants trajectories that were taken from TIMIT DB are shown in black and calculated with our model shown in white. Spectrograms and formant frequencies are shown in the Mel-frequencies range up to 4kHz.

For each of the speech samples the average Root Mean Square Errors (RMSE) for each of the formants is

automatically calculated. In table 1 the results of RMSE calculation for the observed phrases and speakers are shown. The numbers shown in bold cursive letters are the best values of RMSE taken from [38].

Table 1. RMSE values in Hz

Phrase	Speaker	F1	F2	F3	Average
Figure 7	Male 1	92	82	73	82
	Male 2	118	124	132	124
	Male 3	79	84	93	85
Figure 8	Female	115	104	85	101/***108***
Figure 9	Male	170	125	179	158/***285***

As it is seen from figure 7 and table 1 the calculated formant trajectories are very close to those made by hand in TIMIT DB. Moreover, sometimes the calculated trajectories look more preferable (watching the spectrograms maximums) than the hand made ones. It can also be seen that the main deflections of tracks calculated from the TIMIT's tracks are in the regions at the beginning and the end of the spectrogram where the signal is close to zero.

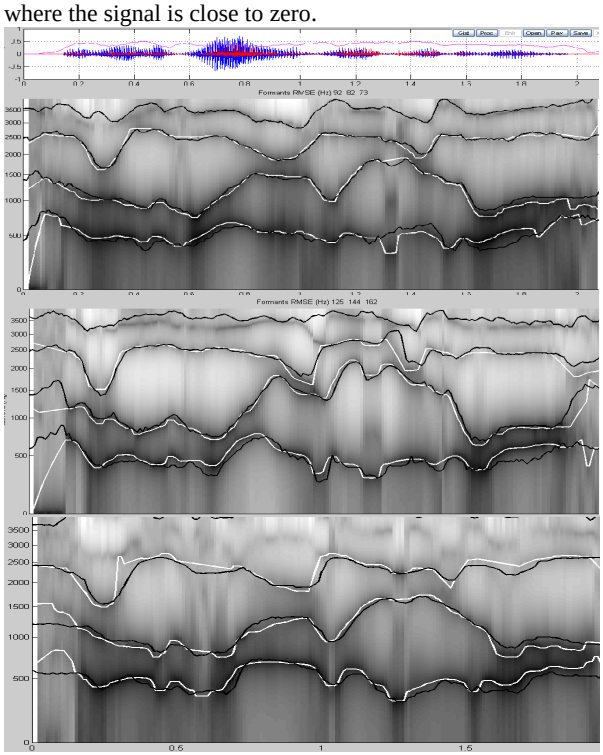

Figure 7 – Formant frequencies tracking result for the phrase "A roll of wire lay near the wall" spoken by three different male speakers.

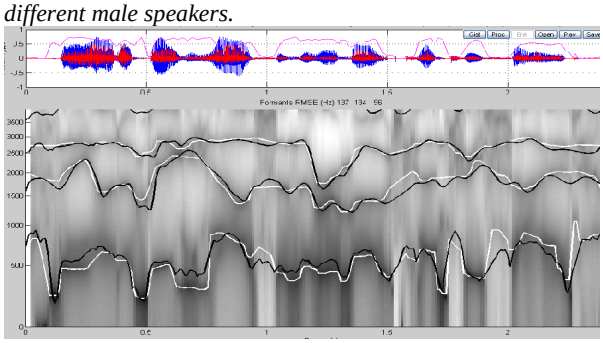

Figure 8 – Formant frequencies tracking result for the phrase "They own a big house in the remote countryside" spoken by female speaker.

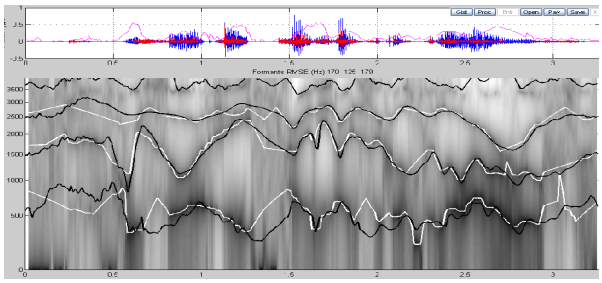

Figure 9 –Formant frequencies tracking result for the phrase "We saw eight tiny icicles below our roof" spoken by male speaker.

Conclusion

Figures 8 and 9 as well as the data from table 1 confirm the efficiency of the method of formant trajectories tracking suggested in this paper. The method works rather well for male voices and less efficiently in the case of a female voice. Further investigations will be directed at increasing the accuracy and robustness of formant analysis both for female and male voices.

References

[1] Potter, R.K., Steinberg, J.C. "Toward the specification of speech," Journl. Acoust. Soc. America. 22, 807-820, 1950.

[2] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[3] Peterson, G. E. "Parameters of vowel quality," J. Speech Hear.Res. 4, 10-29, 1961.

[4] B. Lobanov, *Classification of Russian Vowels Spoken by Different speakers*. Journal Acoust. Soc. America. V. 49, No 2, pp. 606-608, 1971.

[5] B. Lobanov, *On the Classification of Russian Fricatives in C-V syllables for Different Speakers*. Jornal Acoust. Soc. America, N 4 (2), pp. 74-76, 1971.

[6] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[7] P.C. Delattre, A.M. Liberman, F.S. Cooper. "*Acoustic loci and transitional cues for consonants*," *J. Acoust. Soc. Am.* Vol. 27, 1955, pp. 769-773.

[8] B. Lobanov, *On the Acoustic Theory of Coarticulation and Reduction* // Proc. of ICASP-82, Paris, 1982 . – pp. 231-236.

[9] Y. Gao, R. Bakis, J. Huang, B. Zhang, "Multistage coarticulation model combining articulatory, formant, and cepstral features," in *Proc. ICSLP*, vol. 1, 2000, pp. 25–28.

[10] F. Seide, J. Zhou, L. Deng, "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM—MAP decoding and evaluation," in *Proc. ICASSP*, 2003, pp. 748–751.

[11] Yu, D., Deng, L., Acero, A., "Speaker-Adaptive Learning of Resonance Targets in a Hidden Trajectory Model of Speech Coarticulation". Computer Speech and Language. Vol. 27, 2007, pp. 72-87.

[12] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust. Speech Signal Process., vol. 22, pp. 134–141, 1974.

[13] G. Rigoll, "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman-filter," in *Proc. ICASSP*, 1986, pp. 1229–1232.

[14] G. Kopec, “Formant tracking using hidden Markov models and vector quantization,” IEEE Trans. Acoust. Speech Signal Process., vol. 34, pp. 709–729, 1986.
[15] D. Talkin. “Speech formant trajectory estimation using dynamic programming with modulated transition costs” *J. Acoust. Soc. Am.*, S1, 1987, pp. S55.
[16] G. Rigoll, “A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter,” in Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP), 1988, pp. 1229–1232.
[17] G. Rigoll,, “Formant tracking with quasilinearization,” in *Proc. ICASSP*, 1988, pp. 307–310.
[18] D. Broad, F. Clermont, “Formant estimation by linear transformation of the LPC cepstrum,” *J. Acoust. Soc. Amer*., vol. 86, pp. 2013–2017, 1989.
[19] Y. Laprie, M.-O. Berger. Active Model for Regularizing Formant Trajectories. Proc. ICSLP, pp. 815-818, Banf, 1992.
[20] M. Niranjan, I. Cox, “Recursive tracking of formants in speech signals,” in *Proc. ICASSP*, 1994, vol. II, pp. 205–208
[21] D.X. Sun. Robust estimation of spectral center-ofgravity trajectories using mixture spline models. In Proceedings Eurospeech’95, V.1, Madrid, 1995.
[22] L. Welling, H. Ney A Model for Efficient Formant Estimation. Proc. ICASSP, pp. 797-800, Atlanta, 1996.
[23] J.N. Holmes, W.J. Holmes, P.N. Garner. Using Formant Frequencies in Speech Recognition. Proc. EuroSpeech’97, Rhodes – Greece, 1997.
[24] J. Hogberg, “Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients,” KTH-STL Quarterly Progress Rep, Royal Inst. Technol. Stockholm, Sweden, 1997, pp. 41–49.
[25] L. Welling, H. Ney. “Formant tracking for speech recognition,” *IEEE Trans. Speech & Audio Proc.*, Vol. 6, 1998, pp. 36-48.
[26] Lobanov at all. Speaker and Channal-Normalized Set of Formant Parameters for Telephone Speech Recognition. Proceedings of the 6th European Conference on Speech Communication and Technology - EUROSPEECH ’99, Budapest, 1999, pp. 331-334
[27] I. Bazzi, A. Acero, L. Deng, “An expectation-maximization approach for formant tracking using a parameter-free non-linear predictor,” in *Proc. ICASSP*, 2003, pp. 464–467.
[28] L. Deng, I. Bazzi, A. Acero, “Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint,” in *Proc. Eurospeech*, 2003, vol. I, pp. 73–76.
[29] Q. Yan, E. Zavarehei, S. Vaseghi, D. Rentzos, “A formant tracking LP model for speech processing in car/train noise,” in *ICSLP*, Jeju, Korea, Oct. 2004.
[30] Deng, L., Yu, D., Acero, A., “A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech”. Proc. Interspeech 2004.
[31] L. Deng, L.J. Lee, H. Attias, A. Acero. “A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances,” *Proc. ICASSP*, Vol. 1, 2004, pp. 557-560.
[32] Y. Zheng, M. Hasegawa-Johnson, “Formant tracking by mixture state particle filter,” in Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP), 2004, vol. 1, pp. 565–568.
[33] S. Manocha, C.Y. Espy-Wilson, “Knowledge-based formant tracking with confidence measure using dynamic programming,” J. Acoust. Soc. Am., vol. 118, pp. 1930, 2005.
[34] Deng, L., Li, X., Yu, D., Acero, A., “A Hidden Trajectory Model with Bi-Directional Target-Filtering” Proc. ICASP 2005, pp 337-340.
[35] L. Deng, A. Acero, I. Bazzi, “Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint,” IEEE Trans. Audio Speech Lang. Process., vol. 14, pp. 425–434, 2006.
[36] J. Darch, B. Milner, S. Vaseghi, “MAP prediction of formant frequencies and voicin8 class from MFCC vectors in noise,” *Speech Communication*, vol. 48, no. 11, pp. 1556–1572, Nov. 2006.
[37] K. Mustafa, I.C. Bruce, “Robust formant tracking for continuous speech with speaker variability,” IEEE Trans. Audio, Speech and Lang. Process., vol. 14, no. 2, pp. 435–444, 2006.
[38] D. Rudoy, D.N. Spendley, P.J. Wolfe, Conditionally Linear Gaussian Models for Estimating Vocal Tract Resonances, *INTERSPEECH 2007 pp. 526-529*
[39] L. Deng, L.J. Lee, H. Attias, A. Acero, “Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model,” IEEE Trans. Audio Speech Lang. Process., vol. 15, pp. 13–23, 2007.
[40] C. Gl¨aser, M. Heckmann, F. Joublin, C. Goerick, Auditory-based Formant Estimation in Noise using a Probabilistic Framework INTERSPEECH 2008 pp. 2606-2609.
[41] L. Deng, J. Ma, “Spontaneous speech recognition using a statistical coarticulatory model for vocal-tract-resonance dynamics,” *J. Acoust. Soc. Amer.*, vol. 108, pp. 3036–3048, 2000.
[42] B. Strope, A. Alwan, “Robust word recognition using threaded spectral peaks,” *Proc. ICASSP*, 2004, Vol.II, pp. 625-629.
[43] L. Deng, D. Yu, A. Acero. A bi-directional target-filtering model of speech coarticulation and reduction: Two-stage implementation for phonetic recognition,” *IEEE Trans. Speech & Audio Proc.*, Vol. 14, 2006, pp. 256-265.
[44] Whiteside, S. "Sex-specific fundamental and formant frequency patterns in a cross-sectional study". J. Acoust. Soc. Am., 110, 464–478, 2001.
[45] Mezghani, A., O’Shaughnessy, D., Speaker Verification Using a New Representation Based on a Combination of MFCC and Formants, IEEE Canadian Conference on Electrical and Computer Engineering, Saskatoon, SK, May 2005.
[46] Tanabian, M-M., Tierney, P., Zahirazami, B., Automatic speaker recognition with formant trajectory tracking using CART and neural networks, IEEE Canadian Conference on Electrical and Computer Engineering, Saskatoon, SK, May 2005.
[47] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, A. Alwan, “A database of vocal tract resonance trajectories for research in speech processing,” in Proc. IEEE Int. Conf. on Audio, Speech and Signal Process. (ICASSP), 2006, pp. 60–63.
[48] http://www.speech.kth.se/snack/
[49] X. Huang, A. Acero, H.-W. Hon Spoken Language Processing: A Guide to theory, algorithm, and system development – New Jersey: Prentice Hall. – 2001. – 1008 p.