

# Combination of LSF and Pole Based Parameter Interpolation for Model-Based Diphone Concatenation

Karl Schnell, Arild Lacroix

Institute of Applied Physics, Goethe-University Frankfurt  
 Max-von-Laue-Str. 1, D-60438 Frankfurt am Main, Germany  
 schnell@iap.uni-frankfurt.de

## Abstract

For speech generation using small databases, spectral smoothing at the unit joints is necessary and can be realized by an interpolation of model parameters. For that purpose, the LSF are the best choice from the conventional parameter descriptions. This contribution shows how LSF interpolations can be improved using poles as parameters. The problem of the pole assignment between the two pole configurations at the unit joints is solved by pole tracking of an LSF transition. An inspection of the assignments determined by LSF transitions reveals unfavorable cases which can be corrected. A comparison between the LSF and the pole based interpolations shows that the LSF interpolations can be improved by the corrected pole assignments and by the trajectories of the poles. The investigations are performed using a diphone database which is analyzed by an extended LPC model in lattice structure including vocal tract losses.

**Index Terms:** diphone concatenation, spectral smoothing methods, pole assignment

## 1. Introduction

Speech generation is usually performed by a concatenation of speech units. In the case of a large database, selection of units enables direct concatenations yielding high speech quality [1]. In this contribution, a diphone database is used representing a small database. In the case of a small database, spectral smoothing at the unit joints is necessary [2, 3, 4]. For that purpose, a model-based representation of the speech units is suitable since the spectral smoothing can be achieved by an interpolation of the model parameters at the unit joints. To achieve best results, the model parameters should be suitable for interpolation in particular. For models which are based on linear prediction, possible parameter descriptions are reflection coefficients, LAR, LSF etc. Several investigations have shown that the LSF (line spectral frequencies) are well suitable for interpolation and are advantageous to the other parameter descriptions mentioned above, e.g. [5]. However, LSF interpolations yield not always best results for unit concatenation [3, 6]. One potential candidate for the interpolation is also a parameter description by poles due to their meaningful interpretation [3, 7]. One crucial task for that approach is to perform the pole assignment between the two pole configurations [3, 8, 9]. In this contribution, a combination of the pole and LSF description is proposed improving the results in comparison to those obtained by LSF alone. The model used for synthesis is an extended LPC-model in lattice structure.

## 2. Model-based description of diphones

The common lattice filter can be interpreted as a lossless tube model. In this contribution, an extension of that model is used

considering also the losses of the vocal tract. The extended model for synthesis, referred to as the lossy tube model, is depicted in fig. 1. In comparison to the standard lattice filter, the delays  $z^{-1}$  are substituted by lossy delays  $\mathcal{G} = V(z) \cdot z^{-1}$  as shown in fig. 1. The pole-zero system  $V(z)$  models vibrating walls, viscous friction, and heat conduction. Furthermore, the lip termination is realized by the pole-zero model  $\alpha L(z)$  considering losses caused by radiation from the lips. The introduced losses result in a more realistic vocal tract model affecting the bandwidths of the resonances. The lossy tube model is explained in more details in [10]. The

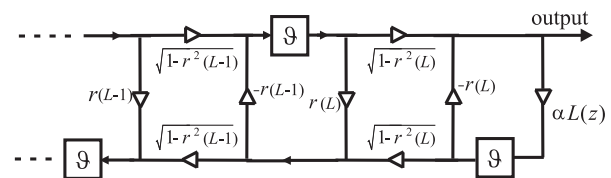


Figure 1: Lossy tube model.

parameters of the tube model to be estimated are the reflection coefficients represented by the vector  $\mathbf{r} = (r(1), r(2), \dots, r(L))^T$  with  $L = 24$ . For the analysis, the diphones are segmented into overlapping frames which are filtered by an adaptive pre-emphasis. The reflection coefficients are estimated from each filtered frame resulting in the vector sequence  $\mathbf{r}_k^{\lambda-\theta}$  which represents the analyzed diphone  $/\lambda - \theta/$ ; the variables  $\lambda$ ,  $\theta$ , and  $\psi$  stand for arbitrary speech sounds with  $\lambda, \theta, \psi \in \{/a/, /z/, /i:/, \dots\}$ . The estimation of the parameter vectors is performed by a minimization of a spectral distance between the magnitude response of the tube model and the filtered diphone segment which is explained in [10, 11]. For the synthesis, the tube model in fig. 1 is controlled by the parameter vectors. The excitation of the model is independent from the analyzed speech. The naturalness of the synthesis depends on the voiced excitation, too. Here, a pitch-modified residual of an individual utterance of the schwa-sound is used for all voiced sounds avoiding unnatural perception like the buzz. The pitch modification algorithm is explained in [12]. In the case of unvoiced fricatives, the excitation is noise-like.

## 3. Model-based diphone concatenation

The diphone concatenation is performed using the corresponding parameter vectors of the diphones. For that purpose, an interpolation of the parameter vectors at the diphone joints is carried out smoothing the spectral discontinuities. For a general notation, the parameter vectors are denoted by  $\mathbf{p}$ . If  $\mathbf{p}_{rh}^{\lambda-\theta}$  is the right-hand parameter vector of the diphone  $/\lambda - \theta/$  and  $\mathbf{p}_{lh}^{\theta-\psi}$  is the left-hand

parameter vector of the diphone  $/\theta-\psi/$ , the concatenation can be described by a transition from the parameter vector  $\mathbf{p}_{rh}^{/\lambda-\theta/}$  to the vector  $\mathbf{p}_{lh}^{/\theta-\psi/}$ . The transition can be described by the vector sequence  $\mathbf{p}_k^{/\lambda-\theta-\psi/}$  for  $k=1\dots N$  with  $\mathbf{p}_1^{/\lambda-\theta-\psi/} = \mathbf{p}_{rh}^{/\lambda-\theta/}$  and  $\mathbf{p}_N^{/\lambda-\theta-\psi/} = \mathbf{p}_{lh}^{/\theta-\psi/}$ . The quality of the resulting diphone concatenation obtained by the parameter transition depends on the parameter description. One problem of assessing the quality of the concatenations is to cover all diphone combinations. Therefore, besides of perceptive tests of synthesized concatenation examples, the use of distortion measures can be helpful to handle a large number of possible diphone concatenations; especially, if constant variables of the concatenation procedure are to be tuned to achieve optimum results. The distortion measure or error  $e$  assesses the smoothness of the transition in the spectral domain. For that purpose, the magnitude responses  $H_k^{/\lambda-\theta-\psi/}$  corresponding to the parameter vectors  $\mathbf{p}_k^{/\lambda-\theta-\psi/}$  of the transition are calculated. The sums of Euclidean distances of subsequent magnitude responses and their derivatives with respect to frequency result in the values  $\varepsilon$  and  $\varepsilon'$ . The geometric mean of  $\varepsilon$  and  $\varepsilon'$  leads to the error  $e$  by

$$e = c\sqrt{\varepsilon \cdot \varepsilon'} \quad (1)$$

$$\text{with } \varepsilon = \sum_{k=1}^{N-1} \sqrt{\sum_i |\tilde{H}_{k+1}(\omega_i) - \tilde{H}_k(\omega_i)|^2},$$

$$\varepsilon' = \sum_{k=1}^{N-1} \sqrt{\sum_i \left| \left( \tilde{H}_{k+1}(\omega_{i+1}) - \tilde{H}_{k+1}(\omega_i) \right) - \left( \tilde{H}_k(\omega_{i+1}) - \tilde{H}_k(\omega_i) \right) \right|^2},$$

and using  $\tilde{H}_k = 20 \cdot \log_{10}(|H_k^{/\lambda-\theta-\psi/}|)$ .  $\omega_i = \pi \cdot i/W$  is the discrete normalized frequency. Here,  $W = 250$  and  $N = 7$  is used.  $c$  is a constant proportional to  $1/((N-1)\pi)$ . To obtain general assessments, the errors  $e$  are calculated for a sufficient number of transitions with different diphone combinations. For each sound representing a voiced sound or a fricative 45 diphones containing this sound are used. The errors of all possible combinations of the diphones are averaged resulting in the error  $\bar{e}$ . The error  $\bar{e}$  represents the mean error of 15500 diphone combinations. The analyzed diphones sampled at 16 kHz are from the German diphone database de1 uttered by a female speaker; for each diphone one candidate exists in the database.

### 3.1. Pole tracking of LSF transitions

The LSF coefficients are usually calculated from the denominator  $A(z)$  of an all-pole model. Since the lossy tube model is an extended version of a standard lattice filter, the LSF  $\gamma_v$  are calculated from a polynomial  $A_v(z)$  considering lossy delays in the tube model. If  $A(z)$  is the polynomial obtained from the reflection coefficients under the assumption of the lossless standard tube model,  $A_v(z)$  is obtained by the substitution  $z \rightarrow z/v$  from  $A(z)$  resulting in  $A_v(z) = A(z/v)$ .  $v$  is a real loss-factor and its value  $v = 0.98$ .

#### 3.1.1. Analysis of LSF transitions

Using poles as model parameters implies difficult tasks: their interpolation and especially the assignment of the poles between the diphone boundary segments. Particularly, the pole assignment is not trivial; one problem is that real and

complex poles can be involved concurrently. To solve this problem, pole matching obtained from LSF transitions is used. For that purpose, a transition  $\tilde{\mathbf{p}}_k^{/\lambda-\theta-\psi/}$  (denoted by a tilde) with relatively many interpolations  $M$  between the two parameter configurations  $\mathbf{p}_{rh}^{/\lambda-\theta/}$  and  $\mathbf{p}_{lh}^{/\theta-\psi/}$  is calculated in the description of LSF; the number of interpolations is denoted by  $M$ . Then, for each interpolated configuration  $k$  the poles  $\tilde{Z}_k^{/\lambda-\theta-\psi/}$  are calculated based on the polynomial  $A_v(z)$ . The vectors of the poles are denoted by  $\mathbf{Z} = (Z(1), Z(2), \dots)^T$ . Since the complex poles exist in conjugated complex pairs, the  $Z(i) = R(i) \cdot e^{j\varphi(i)}$  which are the roots of  $A_v(z)$  with  $0 \leq \varphi \leq \pi$  are sufficient as parameters. The starting and final configuration of the transition are  $\tilde{\mathbf{Z}}_1^{/\lambda-\theta-\psi/} = \mathbf{Z}_{rh}^{/\lambda-\theta/}$  and  $\tilde{\mathbf{Z}}_M^{/\lambda-\theta-\psi/} = \mathbf{Z}_{lh}^{/\theta-\psi/}$ , respectively. Fig. 2 shows an example of the resulting poles of the transition  $\tilde{\mathbf{p}}_k^{/\lambda-\theta-\psi/}$  for  $k=1\dots M$ ; for illustration, the transition is performed with a relatively small number  $M=12$ . For the actual analysis,  $M > 100$  is used yielding a dense pole trace. The pole traces of the transition can be well recognized from

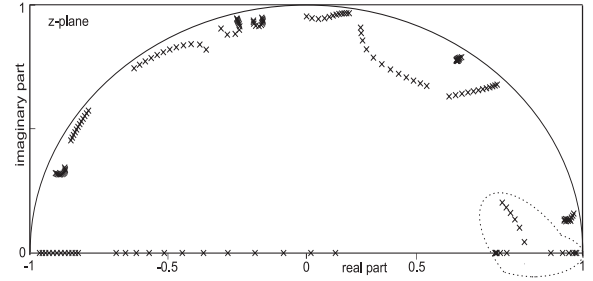


Figure 2: LSF transition; each cross 'x' represents a pole  $\tilde{Z}_k^{/\lambda-\theta-\psi/}(i)$  corresponding to the LSF transition.

fig. 2. An algorithmic determination of the pole traces can be performed by pole tracking with a next-neighbor-strategy. For example, if for the pole  $\tilde{Z}_k^{/\lambda-\theta-\psi/}(i)$  the corresponding pole  $\tilde{Z}_{k+1}^{/\lambda-\theta-\psi/}(j)$  in the next interpolation  $k+1$  is in demand, the one with the smallest distance is chosen

$$\arg \min_j |\tilde{Z}_k^{/\lambda-\theta-\psi/}(i) - \tilde{Z}_{k+1}^{/\lambda-\theta-\psi/}(j)|. \quad (2)$$

In cases of possible pole crossings, also the motion of the poles should be taken into account. In this way, each pole can be connected with a pole of the next configuration representing a pole assignment between two interpolations. The chain of those assignments results in an assignment between the poles of the starting configuration  $\tilde{\mathbf{Z}}_1^{/\lambda-\theta-\psi/} = \mathbf{Z}_{rh}^{/\lambda-\theta/}$  and the poles of the final configuration  $\tilde{\mathbf{Z}}_M^{/\lambda-\theta-\psi/} = \mathbf{Z}_{lh}^{/\theta-\psi/}$ . The resulting assignments can lead to conversions from two real poles to a conjugated complex pole-pair or vice versa, as occurred, for example, by the constellation of fig. 2 at the bottom right which is enclosed by dashed line. The assignments are classified into two groups according to exchanges between real and complex poles. One group  $G$  contains the assignments without exchange and another group  $\bar{G}$  contains the assignments with exchange between real and complex poles. For example, the resulting pole assignment of the enclosed trace in fig. 2 at the bottom right would be classified into group  $\bar{G}$  whereas the assignments from the other pole traces would be classified

into group  $G$ . Additionally to the assignments, also the pole traces are stored for group  $\bar{G}$ .

### 3.2. Transition with poles

The results of the pole tracking of the previous section provide the pole assignment needed for pole transitions. The poles of  $G$  with pole assignments complex to complex and real to real are interpolated by rules. In comparison to that, the pole traces of  $\bar{G}$  implying exchanges between real and complex poles are adopted. For that purpose, the traces of  $\bar{G}$  are only adapted to the new number  $N$  of interpolations which can be easily performed since  $N \ll M$  is valid. In the following, the transition for the poles of group  $G$  is explained. The vectors  $\mathbf{Z}'_k^{\lambda-\theta-\psi'}$  for  $k=1$  and  $k=N$  contain the poles of  $G$  ordered corresponding to their assignments. The rule-based transition is defined separately for the absolute value  $R$  and the angle of the poles described by  $Z_k^{\lambda-\theta-\psi'}(i) = R_k(i) \cdot e^{j\varphi_k(i)}$  for each pole. The angle  $\varphi$  is linearly interpolated. The absolute values are interpolated linearly in the description of the inverse hyperbolic tangent function  $\text{Artanh}(R)$ ; this is performed since the frequency response is more sensitive for values  $R$  close to one than for values significantly smaller than one, similar to the reflection coefficients. Additionally, the absolute values are decreased slightly by the factor  $b$  in the middle of the transition depending on the degree of the frequency shifting  $|\varphi_N(i) - \varphi_1(i)|$ . The reason for this approach is that a strong frequency shifting of resonances with small bandwidths implies unfavorable spectral alterations. The trajectories of the poles  $Z_k^{\lambda-\theta-\psi'}(i) = R_k(i) \cdot e^{j\varphi_k(i)}$  are defined by

$$\varphi_k(i) = (1 - f_k) \cdot \varphi_1(i) + f_k \cdot \varphi_N(i), \quad (3)$$

$$R_k(i) = b(k, i) \tanh((1 - f_k) \text{Artanh}(R_1(i)) + f_k \text{Artanh}(R_N(i)))$$

$$\text{with } b(k, i) = 1 - (1 - 2|f_k - 0.5|) \cdot \beta \cdot |\varphi_N(i) - \varphi_1(i)|;$$

$f_k = (k-1)/(N-1)$  for  $k=1 \dots N$  is an interpolation-variable between zero and one. The constant  $\beta = 0.22$  is chosen in order to minimize  $\bar{\epsilon}$ . The resulting transition contains the same pole matching as the LSF transition, however, with modified pole trajectories for the poles mapping complex to complex or real to real poles. Fig. 3 shows an example of

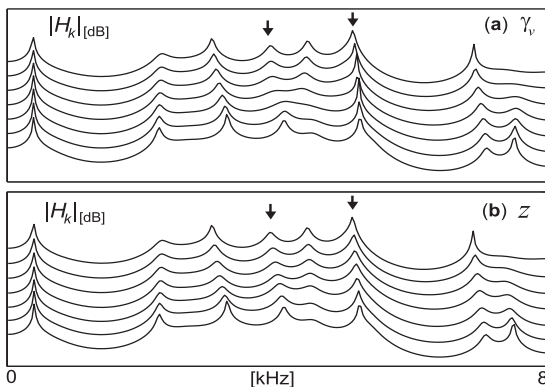


Figure 3: Transition of magnitude responses for the diphone concatenation within the sound /j/: (a) LSF parameters and (b) pole based parameter description.

concatenation by LSF transition and the transition with pole trajectories. The pole bandwidths of the interpolated

configurations by LSF transition often deviate from a more direct trajectory as expected, which can be seen from the marked resonance movements in fig. 3. In comparison to the LSF transitions, the rule-based pole trajectories lead to correct bandwidths.

#### 3.2.1. Modification of pole matching

Not in all cases, the matching obtained from the LSF transition is the best choice. The inspection of transitions with large error  $e$  reveals also unfavorable pole assignments by LSF transitions. One class of relevant cases can be described by a typical pole constellation which is exemplarily illustrated in the following case study with two complex conjugated poles shown in fig. 4. The poles of the starting configuration  $\mathbf{Z}^o = [Z^o(1), Z^o(2)]$  and the poles of the final configuration

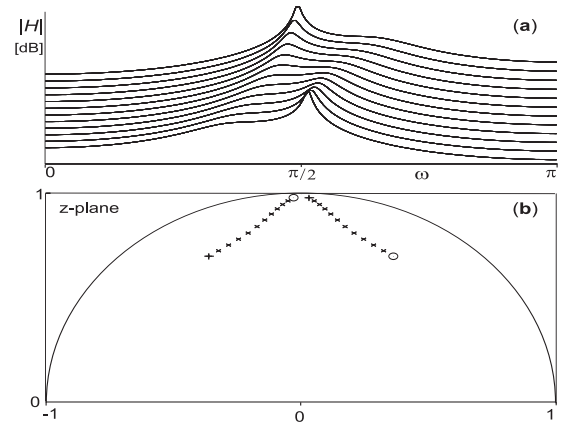


Figure 4: (a) magnitude responses and (b) poles of LSF transition. The poles of the starting configuration  $\mathbf{Z}^o$  are denoted by 'o' whereas the poles of the final configuration  $\mathbf{Z}^+$  are denoted by '+'.

$\mathbf{Z}^+ = [Z^+(1), Z^+(2)]$  are ordered according to their angles with  $\varphi(i) < \varphi(i+1)$ . The case of false assignment occurs, if a pair of poles  $Z^+(i), Z^o(i+1)$  close to the unit circle is surrounded by the poles  $Z^o(i), Z^+(i+1)$  positioned more inside of the unit circle. The membership of the poles to the starting or final configuration alternates  $Z^o(i), Z^+(i), Z^o(i+1), Z^+(i+1)$ , if the poles are ordered according to their angles. The corresponding poles of the transition between  $\mathbf{Z}^o$  and  $\mathbf{Z}^+$  in the description of LSF reveal the assignment  $Z^o(i) \rightarrow Z^+(i), Z^o(i+1) \rightarrow Z^+(i+1)$  by the LSF transition. From the corresponding magnitude responses, it can be seen that the transition is inadequate, since the two high-Q resonances are not linked. The LSF transition has assigned the poles according to the angles. However, in this case an assignment according to the absolute values would be advantageous with  $Z^o(i) \rightarrow Z^+(i+1), Z^o(i+1) \rightarrow Z^+(i)$ .

The described case can be detected by an automatic inspection of the assignments of all possible diphone combinations. The test can be implemented with the aid of thresholds for the absolute values of  $Z(i), Z(i+1)$  representing the analyzed pole pairs. If the test is positive, the assignment is altered to an assignment according to the absolute values of the poles. Fig. 5 gives an example for a detected case and its modification. Fig. 5(a) and (c) represent the diphone transition with LSF. The unfavorable pole

assignment is corrected in the transition with poles, shown in fig. 5(b) and (d). ‘o’ and ‘+’ are the starting and final pole positions, respectively. The arrows indicate the pole configuration which is changed in the manner that the poles with the absolute values near to one are assigned. The pole transition from complex poles to real poles between -1 and -0.7 is originally caused by LSF and is adopted for the pole transition. It can be seen that the modified pole matching improves the transition by the pole trajectories and especially by the modified pole matching. The improvement by modification of the pole matching is effective in ca. 500 of the 15500 combinations.

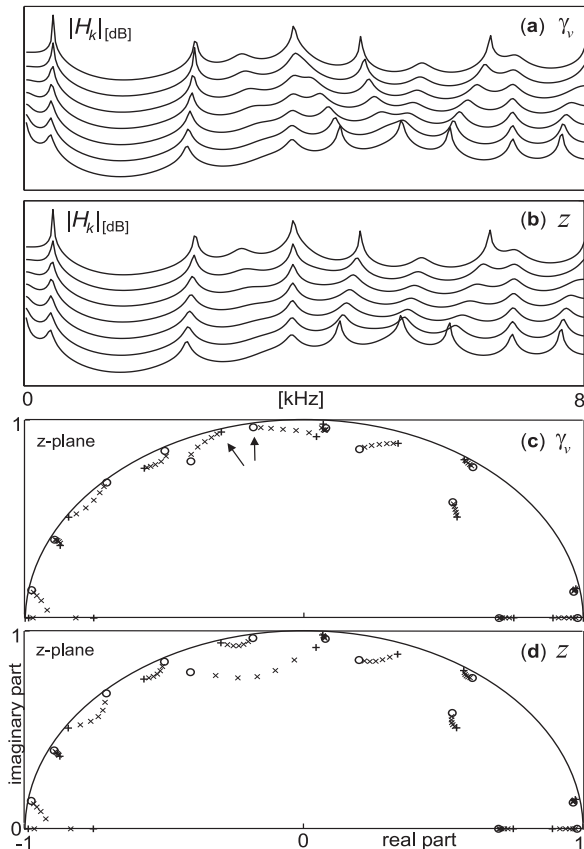


Figure 5: Diphone concatenation: magnitude responses of transition (a) with LSF and (b) with poles and modification of assignment. The corresponding poles of (a) and (b) are in (c) and (d), respectively.

### 3.2.2. Comparison of parameter descriptions

Table 1 shows the averaged errors  $\bar{e}$  depending on the type of model parameter for the diphone transition. Also the

Table 1. Error  $\bar{e}$  depending on parameter description: Reflection coefficients  $r$ , LAR, LSF  $\gamma_v$ , and poles  $Z$ .

	$r$	LAR	$\gamma_v$	$Z$
$\bar{e}$	1.97	2.21	1.42	1.24

errors are listed obtained by transitions with log area ratios and reflection coefficients; these are significantly larger than the error obtained by LSF, which is in coincidence with [5]. Applying different nonlinear functions to the reflection coefficients or areas doesn't change this situation. In comparison to conventional coefficients, the use of the pole

transitions, as described in section 3.2, needs an analysis of the diphone database in advance; however, this analysis has to be performed only once. The averaged improvement of the LSF transitions by the poles is about 13%.

## 4. Conclusions

For the concatenation of model-based diphones, a smoothing method is proposed based on a combination of LSF and pole interpolations. The problem of the pole assignment is solved by an adoption of assignments obtained from LSF transitions. The analysis of the assignments obtained by LSF transitions reveals unfavorable cases which can be detected and corrected. The investigations show that the concatenation results by LSF alone can be improved both by modifications of the pole trajectories and the pole assignments. Additionally to the improved concatenation results, the pole based description gives more insights into the parameter transitions due to their meaningful interpretation in comparison to LSF.

## 5. References

- [1] Hunt, A.J. and Black, A.W., "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database", Proc. ICASSP'96, Atlanta 1996, pp.373-376.
- [2] Shadle, C. H. and Atal, B. S., "Speech synthesis by linear interpolation of spectral parameters between dyad boundaries", J. Acoust. Soc. Amer., Vol. 64, 1979, pp. 1325-1332.
- [3] Chappell, D.T. and Hansen, J.H.L., "A comparison of spectral smoothing methods for segment concatenation based speech synthesis", Speech Communication (36), pp. 343-374, 2002.
- [4] Dutoit, T., An introduction to text-to-speech synthesis. Dordrecht: Kluwer Academic Publishers, 1997.
- [5] Paliwal, K. K., "Interpolation properties of linear prediction parametric representations", Proc. EUROSPEECH'95, Madrid, 1995, pp. 1029-1032.
- [6] Pfitzinger, H. R., "DFW-based Spectral Smoothing for Concatenative Speech Synthesis", Proc. INTERSPEECH'04, Jeju Korea, 2004.
- [7] Pearson, S., Kibre, N., and Niedzielski, N., "A Synthesis Method Based on Concatenation of Demisyllables and a Residual Excited Vocal Tract Model", Proc. ICSLP'98, Sydney 1998, pp. 2739-2742
- [8] Goncharoff, V. and Kaine-Krolak, M., "Interpolation of LPC Spectra via Pole Shifting", Proc. ICASSP'95, Detroit 1995, pp. 780-783.
- [9] Turnbull, J.M., Sapeluk A.T., and Damper, R.I., "A New Method of Pole-Tracking with Application to Vowel and Semivowel Recognition", Proc. ICASSP'89, Glasgow 1989, pp. 568-571.
- [10] Schnell, K. and A. Lacroix, A., "Analysis of lossy vocal tract models for speech production", in Proc. INTERSPEECH'03, Geneva 2003, pp. 2369-2372.
- [11] Schnell, K. and A. Lacroix, A., "Model Based Analysis of a Diphone Database for Improved Unit Concatenation", Proc. INTERSPEECH'05, Lisbon Portugal 2005, pp. 2605-2608.
- [12] Schnell, K., "Pitch Modification of Speech Residual Based on Parameterized Glottal Flow with Consideration of Approximation Error", Proc. ICASSP'06, Toulouse 2006, pp. 737-740.