Adaptation of cepstral coefficients for acoustic-to-articulatory inversion

Julie Busset, Yves Laprie

Adaptation of cepstral coefficients for acoustic-to-articulatory inversion

Julie Busset , Yves Laprie

[1]LORIA CNRS UMR7503
615, rue du Jardin Botanique 54600 Villers-lès-Nancy, France

{Julie.Busset,Yves.Laprie}@loria.fr

Abstract. *Acoustic-to-articulatory inversion of speech signals via an analysis-by-synthesis method requires the comparison of natural and synthetic speech spectra either indirectly via formant frequencies, or directly via cepstral coefficients. This paper investigates several strategies of cepstral adaptation (affine transformation of cepstral coefficients, bilinear or piecewise linear frequency warping) when X-ray images of the speaker's vocal tract are available. These images enable the articulatory synthesis of a speech signal which fits the natural signal at best. It is thus possible to investigate the behavior of several cepstral adaptation procedures in order to select the best method, i.e. that which minimizes the deviation between synthetic and natural spectra. Our results show that the affine cepstral adaptation tends to flatten the spectral peaks, i.e. formants. Frequency warping techniques are thus more efficient all the more they can be supplemented by taking into account the spectral tilt.*

1. Introduction

Our approach of acoustic-to-articulatory inversion (Ouni and Laprie, 2005; Potard and Laprie, 2009) comes within the scope of the analysis-by-synthesis paradigm. The general idea is to compare spectra of natural speech and those synthesized via an articulatory synthesizer. Its main strength is linked to the fact that the link between vocal tract shapes and acoustics is explicitly modeled via the acoustic simulation. This enables the exploration of acoustic consequences of any articulatory modification.

Until now the comparison carried out in our inversion algorithm was about formants because formant frequencies of synthetic speech correctly render those of natural speech. However, formants of natural speech cannot be extracted easily essentially because of the interaction with the excitation signal. This thus often results in errors in the process of inversion.

This works is about the use of cepstral vectors as input of the inversion process. Unlike formants whose frequencies can be easily compared between natural and synthetic speech, cepstral coefficients of natural speech cannot be compared with those of articulatorilly synthesized speech for the following reasons: (i) the source signal is not approximated accurately, or even not taken into account, in synthetic spectra, (ii) the mismatch between the geometries of the speaker's vocal tract and of the articulatory model, (iii) the mismatch between the real acoustics and that given by the acoustic simulation. The first

objective of this work is to enable a direct comparison of cepstral vectors for both types of speech. This issue is generally answered by using lifters (Meyer et al., 1991) designed to enhance the contribution of formants. However, many X-ray articulatory films provide both the acoustic speech signal and the corresponding temporal evolution of the vocal tract shape.

This corresponds to the favorable situation where synthetic cepstral coefficients can be calculated via articulatory synthesis and compared to natural cepstral coefficients. We thus propose to exploit X-ray data to perform the direct adaptation of natural and synthetic cepstral coefficients. In order to investigate the potential of this adaptation it is even possible to derive the speaker's articulatory model from X-ray images in order to reduce the mismatch corresponding to using a generic articulatory model derived from X-ray images of another speaker.

In the first section, we describe the construction of the articulatory model based on Xray images. Then, the methods of cepstral adaptation are presented and evaluated. Finally, we draw conclusions about the way of accessing the cepstral codebook used to represent the articulatory-to-acoustic mapping.

2. Construction of the articulatory model

The objective of an articulatory model is to approximate the vocal tract shape with a small number of parameters corresponding to deformations modes. Here, a second objective is to design a model which fits our subject as well as possible. This explains why we do not use the model of Maeda (1979) we previously used for inversion. Unlike the model evaluated in (Laprie and Busset, 2011), here the conciseness is a crucial objective since the inversion algorithm explores the articulatory space. This explains why the number of deformation modes has been limited to seven and why the model construction differs from that reported in (Laprie and Busset, 2011) even if the same images and articulatory contours were used.

First, the jaw movement given by the rotation and translation of the mandible has been analyzed with PCA. Only the first component which explains 61% of the variance has been kept. Unlike other approaches this first linear component controls both the rotation and the translation of the mandible. Then, the movement of the mandible is subtracted from the tongue contour. An adaptive polar grid, whose center is attached to the mandible, and covering the tongue from the root to the apex has been utilized to get a vector of points representing the tongue contour. PCA was applied to the vectors of tongue points and four components were kept representing 97.6% of the variance (Tab. 1 gives the variances explained by the first six components).

Table 1. *Variance of tongue contours explained by the PCA components*

# of components	1	2	3	4	5	6
Percentage	53.26	32.19	7.53	4.62	0.96	0.56
Total	53.26	85.45	92.98	97.60	98.56	99.12

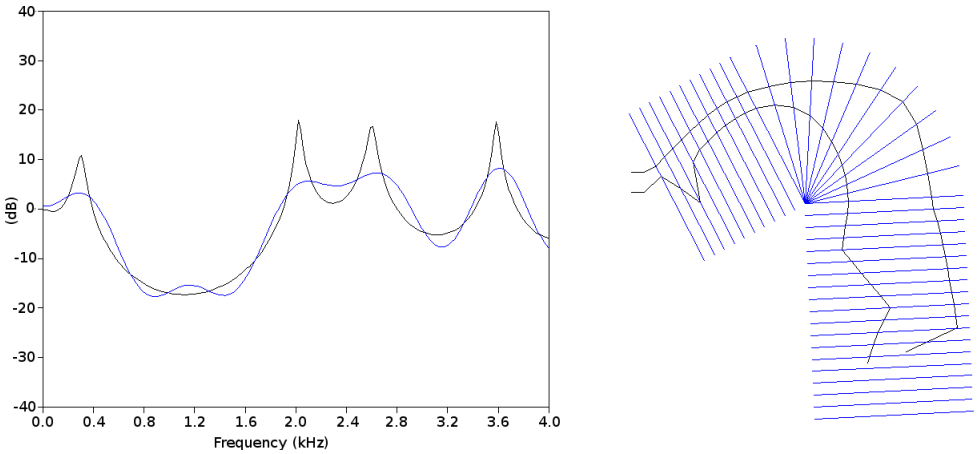

Figure 1. Simplified articulatory model and resulting synthetic spectrum. The spectrum whith sharp peaks is the frequency simulation and the smooth curve is the cepstrally smoothed spectrum with 30 cepstral coefficients.

The lips are represented by the contours of the upper and lower lips. Two parameters are computed: lip height and protrusion. The lip height is the distance between the two contours where the lip aperture is minimal and the protrusion is given by the abscissa of the midpoint of the segment corresponding to the lip aperture. Although the lips are controlled by only two parameters, PCA is applied in order to retain only one parameter. The first component explained 90.19% of the variance.

The epiglottis contour and the larynx are considered as only one articulator. More precisely the epiglottis is given by its contour and the larynx by two points corresponding to its extremities. The percentage of variance explained by the first PCA component is 47.41% and 36.85% for the second. However, only the first component is retained in the global model in order to reduce the number of components.

Since we were not dealing with nasalized sounds we did not incorporate any component rendering the velum movement. The global articulatory model thus consists of seven parameters, one for the jaw, four for the tongue, one for the lips and one for the epiglottis and larynx. The palate contour is fixed and has been delineated in one reference image.

The third step is to calculate the area function corresponding to the 2D shape of the vocal tract. The area function represents the cross sectional area of each section from the glottis to the lips. Most of the sagittal to area algorithms are based on the results of Heinz and Stevens (1965), where the cross-sectional area of one section is defined by the equation $A = \alpha d^{\beta}$ where d is the midsagittal distance, α and β depending of the position along the vocal tract. In our case α and β coefficients used are those defined by Soquet et al. (2002) and the midsagittal distances are determined by applying a semi-polar grid on the midsagittal contours. For each frame, a transfer function associated to the area function is calculated via the acoustic simulation provided with the model of Maeda (1990).

3. Adaptation of cepstral coefficients

The compensation of the mismatch between cepstral coefficients of synthetic and natural speech is very often addressed by using cepstral lifters which attenuate the contribution of first and last coefficients (Meyer et al., 1991). Indeed, the first cepstral coefficients roughly correspond to the spectral tilt and the last coefficients to the harmonics. However, a closer examination of natural speech spectra and those produced by the articulatory synthesizer shows that spectral peaks are also slightly shifted. Liftering techniques thus cannot compensate these deeper modifications. We thus investigated two kinds of transformations: (i) affine transformation of cepstral coefficients, (ii) warping frequency via bilinear transformation or piece-wise linear scaling.

3.1. Linear mapping of cepstral coefficients

The comparison between real and synthetic data has been studied by Mokhtari et al. (2004) in a very similar situation since MRI images and speech signals were available for the same speaker. Mokhtari and his colleagues used linear prediction inversion and compensated formant frequencies and bandwidths via an affine transformation so as to guarantee a better fitting between real and inverted area functions. The coefficients of the affine transformations, one for each formant frequency and bandwidth, were derived from a set of five vowels.

Similarly, we are considering affine transformations of cepstral coefficients to bring cepstral coefficients of natural and synthetic speech closer. We are using linear cepstral coefficients and not MFCC because F2 and F3 often correspond to different acoustic cavities in the vocal tract and we want to preserve them as independent spectral peaks. The linear regression is performed on each cepstral coefficient separately. Hence, each synthetic coefficient is approximated by an affine transformation of the coefficient computed on the real speech signal. For the n^{th} cepstral coefficient, c'_n, the synthetic coefficient is given by:

$$c'_n \approx a_n.c_n + b_n \tag{1}$$

where c_n is the coefficient from the real speech signal. Coefficients a_n and b_n are found by minimizing the error E_n:

$$E_n = \sum_k ||c'_{nk} - (a_n.c_{nk} + b_n)||^2 \tag{2}$$

where n is the index of the coefficient and k the index of the vowel shape used.

Since we are dealing primarily with vowels and because the acoustic simulation is also more precise for vowels we only considered X-ray images corresponding to vowels. For the four X-ray films available for the subject this represents 137 images. The synthetic cepstral coefficients are derived from the spectrum calculated with the acoustic simulation in the frequency domain and natural cepstral coefficients are computed on a 32 ms Hamming windowed signal. The adaptation procedure was applied on the first thirty cepstral coefficients except the very first coefficient which is relative to the signal energy.

The error E_n is minimized via a least square method over the set of the 137 vowels of the corpus. As it can be seen on Fig. 2 this adaptation behaves all the more correctly since the peaks of natural spectra fit those of synthetic spectra. Unfortunately even if

the articulatory model has been constructed specifically for the speaker who uttered the speech signal a perfect fitting cannot be achieved and the minimization thus tends to over-smooth the synthetic spectra to put them closer to the natural spectra on average. More precisely, the adaptation decreases the distance between synthetic and natural spectra from 9.57 dB to 5.54 dB when 30 coefficients are adapted and 5.73 dB when only the first two coefficients are adapted. The benefit of adapting all the coefficients instead of the first two is small, and all the smaller since this gain is obtained to the detriment of the formantic structure of spectra.

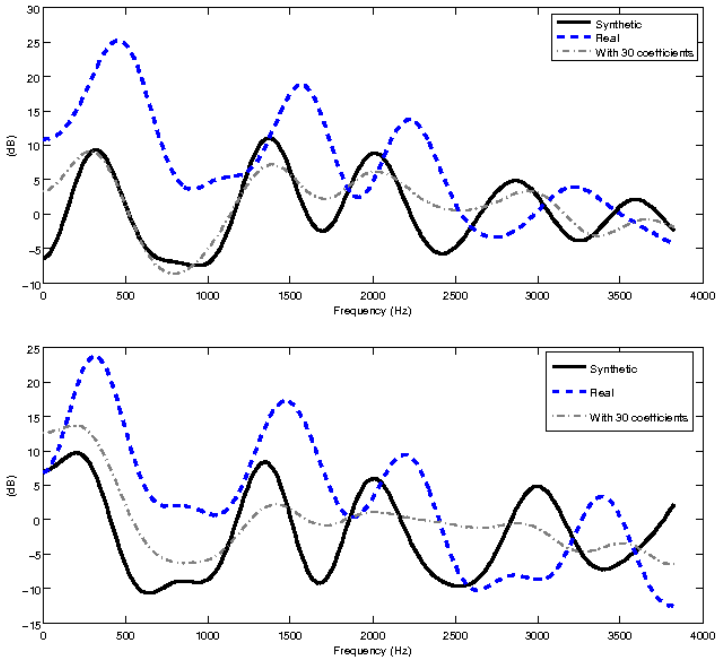

Figure 2. Natural and synthetic cepstrally smoothed spectra: dashed blue line (natural speech), black solid line (synthetic spectrum), gray dashed line (real spectrum adapted to fit the synthetic spectrum). All spectra are calculated with 30 cepstral coefficients.

Fig. 2 shows the result of the adaptation for two examples. The first (Fig. 2 top) corresponds to a successful adaptation. This means that the real spectrum after adaptation is close to the synthetic spectrum. In this case the distance between cepstral coefficients is relevant. On the contrary, the second example (Fig. 2 bottom) shows that the adaptation fails to put the two spectra close together. The adaptation flattens the spectrum and the peaks corresponding to formants which thus partially disappear. Actually the effect of flattening is higher when all the cepstral coefficients are adapted.

Fig. 3 shows the effect of adapting 2 coefficients instead of the 30 coefficients. It clearly appears that even if the peak frequencies are shifted correctly when using 30 coefficients the formantic structure is degraded which compromises the codebook exploration during inversion.

The adaptation is thus relevant when it is applied on the very first cepstral coefficients to capture the spectral tilt of speech but counter-productive if it is applied on all

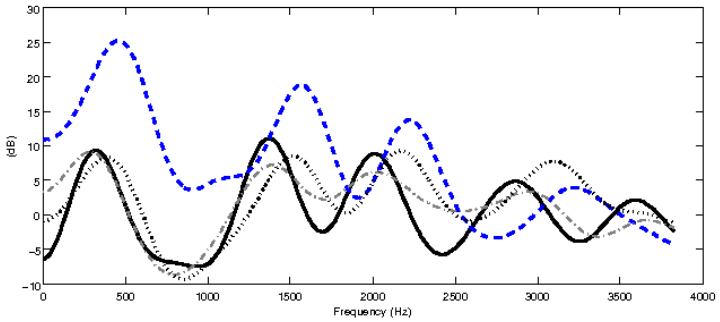

Figure 3. Natural and synthetic cepstrally smoothed spectra: dashed blue line (natural speech), black solid line (synthetic spectrum), gray dashed line (all coefficients adapted), dark gray dotted line (first two coefficients adapted only).

the cepstral coefficients. In the following two sections we investigated frequency warping intended to shift spectral peaks so as to get a better peak fitting between the natural and synthetic spectra without flattening the spectra.

3.2. Frequency warping

Usually frequency warping is used in automatic speech recognition to carry out speaker adaptation. In our case the articulatory model has been constructed from images of the speaker whose speech is inverted. Frequency warping is thus intended to compensate for residual frequency deviations due to the model mismatch or the calculation of the vocal tract midline.

The bilinear transform is a classical tool used in automatic speech recognition to perform vocal tract length normalization. It gives the new frequency variable z_{new} according to the following expression:

$$z_{new} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \qquad -1 < \alpha < 1$$

where α is the parameter of the warping. α affects the whole frequency scale (Oppenheim and Johnson, 1972):

$$\omega_{new} = \omega + 2tan^{-1}\frac{\alpha sin\omega}{1 - \alpha cos\omega}$$

Here, the adaptation consists in choosing α so as to reduce the deviation between peaks of natural and synthetic spectra.

A second solution to implement a frequency warping is to use a piecewise linear scaling. Compared to the bilinear transform which affects the whole frequency domain the piecewise linear scaling can be easily focused on the spectral domain corresponding on F1-F3 formants, i.e. the most important formants from the acoustic-to-articulatory inversion point of view. However, this requires at least two parameters to be adjusted (one to decompose the frequency domain and one to set the warping level). Hence, we have chosen the bilinear transform which only requires one parameter.

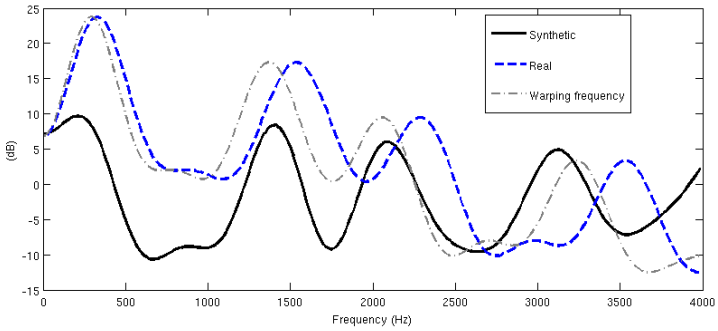

Figure 4. Natural and synthetic cepstrally smoothed spectra: dashed blue line (natural speech), black solid line (synthetic spectrum), gray dashed line (real spectrum after bilinear frequency warping).

Fig. 4 shows the effect of the frequency warping after optimization on the 137 vowels. It can be seen that the formant structure is well preserved which guarantees that the relevant region in the articulatory domain would be explored during inversion.

4. Discussion and concluding remarks

The very first conclusion concerns the precision which can be reached by articulatory copy synthesis. Here the situation is reasonably favorable since the articulatory model fits the geometry of the speaker's vocal tract well. The fitting could be improved by increasing the number of deformation modes used in the model, or by using a 3D model but it would not be realistic in terms of applications. However, despite this favorable situation there remains a non-negligible deviation between synthetic and natural spectra, and more precisely a frequency shift between spectral peaks. The inversion process generally compensates for this acoustic deviation by introducing an articulatory bias in the articulatory trajectories recovered. Since the overall geometry is correct one origin could be the midline used to decompose the vocal tract into elementary uniform tubes.

The experiments presented above show that the frequency warping is more appropriate to adapt the natural speech signal than the affine transform of cepstral coefficients. It is possible to combine both methods by keeping the affine adaptation for the very first cepstral coefficients to capture the spectral tilt without destroying the formantic structure and by warping the frequency scale to get a better fitting between spectral peaks.

The second conclusion concerns the spectral tilt. Modal phonation corresponds to a glottal spectra tilt of -12 dB/octave and lip radiation to a tilt of +6 dB/octave thus resulting in global spectral tilt of around -6dB/octave. However, the glottal tilt is not constant and the affine adaptation of cepstral coefficients is sometimes unable to compensate for it. Hence, the spectral distance could remain substantial even if the formantic structure is fairly well approximated. The correlation between the adapted and synthetic spectra is thus probably more relevant to compare synthetic and natural speech than the euclidian distance.

These results will be exploited to access the articulatory codebook used to represent the articulatory-to-acoustic mapping. Each entry of the codebook associates one

vector of articulatory parameters with one vector of cepstral coefficients. The construction strategy is very similar to that we have utilized with formants and consists of subdividing the articulatory domain until the mapping becomes sufficiently linear (Potard and Laprie, 2007). The codebook is thus organized as a hierarchy of hypercuboids and the inversion process consists of selecting the articulatory cuboids providing the smallest acoustic distance.

References

Heinz, J. M. and Stevens, K. N. On the relations between lateral cineradiographs, area functions and acoustic spectra of speech. In *Proceedings of the 5th International Congress on Acoustics*, page A44., 1965.

Laprie, Y. and Busset, J. A curvilinear tongue articulatory model. In *The Ninth International Seminar on Speech Production - ISSP'11*, Canada, Montreal, 2011.

Maeda, S. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

Maeda, S. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

Meyer, P., Schroeter, J., and Sondhi, M. M. Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks. *IEEE Trans. ASSP*, 39(7):1493–1502, 1991.

Mokhtari, P., Kitamura, T., Takemoto, H., and Honda, K. Evaluation of an lp-based method of inversion using mri-based vocal-tract area functions. In *Autumn Meeting of the Acoustical Society of Japan*, pages 237—238, Okinawa, Japan, 2004.

Oppenheim, A. V. and Johnson, D. H. Discrete representation of signals. *IEEE Trans. on Speech, and Audio Processing*, 60(6):681–691, January 1972.

Ouni, S. and Laprie, Y. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 118 (1):444–460, 2005.

Potard, B. and Laprie, Y. Compact representations of the articulatory-to-acoustic mapping. In *Interspeech, Antwerp*, August 2007.

Potard, B. and Laprie, Y. A robust variational method for the acoustic-to-articulatory problem. In *10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009*, United Kingdom, Brighton, 2009.

Soquet, A., Lecuit, V., Metens, T., and Demolin, D. Mid-sagittal cut to area function tranformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication*, 36(3-4):169–180, March 2002.