Acoustic-to-articulatory inversion by analysis-by-synthesis using cepstral coefficients

Julie Busset, Yves Laprie

Acoustic-to-articulatory inversion by analysis-by-synthesis using cepstral coefficients

Julie Busset, Yves Laprie

LORIA/CNRS, UMR 7503
615, rue du Jardin Botanique 54600 Villers-lès-Nancy, France
Julie.Busset,Yves.Laprie@loria.fr

Abstract

This paper deals with acoustic to articulatory inversion of speech by using an analysis by synthesis approach. We used old X-ray films of one speaker to (i) the develop a linear articulatory model presenting a small geometric mismatch with the subject's vocal tract mid sagittal images (ii) and design an adaptation procedure of cepstral vectors used as input data. The adaptation exploits the bilinear transform to warp the frequency scale in order to compensate for deviation between synthetic and natural speech. This enables the comparison of natural speech against synthetic speech without using cepstral liftering. A codebook is used to represent the forward articulatory to acoustic mapping and we designed a loose matching algorithm using spectral peaks to access it. This algorithm, based on dynamic programming, allows some peaks in either synthetic spectra (stored in the codebook) or natural spectra (to be inverted) to be omitted. Quadratic programming is used to improve the acoustic proximity near each good candidate found during codebook exploration. The inversion has been tested on speech signals corresponding to the X-ray films. It achieves a very good geometric precision of 1.5 mm over the whole tongue shape unlike similar works evaluating the error at 3 or 4 points corresponding to sensors located at the front of the tongue.

INTRODUCTION

Medical imaging techniques applied to the study of speech production have shown a spectacular improvement since the appearance of magnetic resonance imaging, which offers either a good three dimensional geometrical precision, or lately a good temporal resolution (Narayanan *et al.*, 2011). Additionaly, there exists a quantity of old Xray data which are now available (Sock *et al.*, 2011) and offer a good temporal resolution of the mid-sagittal view of the vocal tract. Together with numerical acoustic simulations it is thus possible to investigate speech production or to tackle acoustic-to-articulatory mapping with an analysis-by-synthesis algorithm.

Even if this approach still requires improvements in the acoustic simulation of consonants and acquisition of articulatory data we are convinced that it presents a very good potential for the future. It is especially true when few data or non homogeneous data are available for training like articulatory rehabilitation of hard of hearing patients or foreign language learning. Indeed, in both cases the sounds produced by the subjects may be too far from the articulatory data used for training, and consequently may not be correctly inverted by statistical techniques.

We thus developed an analysis-by-synthesis method (Ouni and Laprie, 2005) where inversion is performed by computing parameters of an articulatory model to match acoustic features computed from natural speech. The underlying articulatory model is built from X-ray images of the sagittal view of the vocal tract. Seven parameters control the sagittal shape of the vocal tract ; one for the jaw opening, four for the shape and the position of the tongue, one for the opening and protrusion of lips and one for the height of the larynx.

In previous works (Potard and Laprie, 2009) we used frequencies of the first three formants as input data. However, formants often cannot be extracted from speech reliably, which thus causes errors in inversion.

This work is dedicated to the use of cepstral vectors as input data. This changes the inversion problem in depth at least for two reasons. First, the cepstral vector has more coordinates than the number of articulatory parameters with the risk of finding no inverse solution at all. Second, the effect of the excitation source and of the mismatch between the speaker's vocal tract and the articulatory model must be taken into account explicitly to compare natural spectra against those generated by the articulatory synthesizer.

From a practical point of view, the main difficulties thus are the comparison of natural and synthetic speech on the one hand, and the access to the codebook on the other hand. This paper mainly focuses on these two points.

Unlike automatic speech recognition where mel frequency cepstral coefficients are used, we here use linear cepstra because the spectral peaks corresponding to formants are preserved. We are currently using a window of 32 ms and 30 coefficient vectors which remove the contribution of harmonics. Speech is pre-emphasized and Hamming windowed.

Experiments presented in this paper have been carried out on four X-ray films the DocVacim database (Sock *et al.*, 2011) recorded by the same subject.

ACOUSTIC-TO-ARTICULATORY INVERSION WITH CEPSTRAL COEFFICIENTS

Local solving with cepstra

The articulatory-to-acoustic mapping is represented in the form of a codebook, i.e. a collection of small articulatory regions where the mapping can be linearized with little loss of precision. We will describe the construction of the codebook and the way of finding regions corresponding to the input cepstral vector in the next sections. Here, we make the hypothesis that an articulatory region, i.e. a parallelepiped box (also called cuboid) of the codebook, may correspond to a cepstral input vector and we want to find the inverse solutions if they exist. Let be c the input cepstral vector and f the local articulatory-to-acoustic mapping. We are searching for the articulatory vector x such that:

$$f(x) = c \tag{1}$$

The equation (1) is an overdetermined system since there are more equations, i.e. one for each of cepstral coefficients, than unknowns, i.e. one for each articulatory parameter. Indeed, the articulatory model uses seven parameters to describe the vocal tract shape and the linear cepstral vector has 30 coefficients. Generally, there may be no exact solution and we thus resort to a least squares fitting method, which minimizes the distance between the target cepstral vector and the acoustic image of the articulatory synthesis:

$$\min_x ||f(x) - c||^2 \tag{2}$$

In addition, the solutions must belong to the parallelipiped considered, what gives rise to two constraints:

$$x < P_0 + s/2 \tag{3}$$

$$x > P_0 - s/2 \tag{4}$$

where s is the size vector, and P_0 the center of the parallelipided. In each parallelepiped f is linearized and thus $f(x)$ is approximated by Ax where A is a matrix. The acoustic distance to be minimized is thus:

$$\begin{aligned} D(x) &= (c - Ax)^\top (c - Ax) \\ &= c^\top c + x^\top \underbrace{A^\top A}_{H} x - (\underbrace{c^\top A}_{q} x + x^\top \underbrace{A^\top c}_{q^\top}) \\ &= x^\top H x - 2q^\top x + r \end{aligned} \tag{5}$$

Equation (2) under the contraints defined by Eq. (4) is solved by using a quadratic programming method, i.e. the Goldfarb and Idnani algorithm (Goldfarb and Idnani, 1983) which minimizes a function f defined by: $D(x) = q^\top x + \frac{1}{2} x^\top H x$ with the constraints: $s(x) = C^\top x - B \geq 0$ where C and B are derived from P_0 and s.

The difficulty of inversion consists of finding relevant articulatory regions, i.e. cuboids of the codebook, where the resolution of the inverse mapping can be applied, and of preparing cepstral data, i.e. the c vector.

We will describe these aspects in the next sections.

Codebook construction and search

The articulatory codebook is a collection of articulatory regions, more precisely cuboids, which cover the articulatory space and where the articulatory-to-acoustic mapping can be linearized with a reasonable precision. The codebook is constructed by dividing recursively the articulatory space into cuboids where the deviation between the articulatory-to-acoustic mapping and a linear mapping is sufficiently small. The construction algorithm is very similar to that used with formants (Potard and Laprie, 2007) except that the linearity is evaluated on the cepstral vectors directly.

The main issue with codebooks is to find out all relevant entries considering one acoustic vector to be inverted. This is simple with formant frequencies as input data since all the parallelepipeds forming the codebook can be indexed with respect to the first three formant frequencies. This solution is no longer possible with cepstral coefficients since one cepstral coefficient alone cannot be used to find all relevant regions. We will describe the adaptation of cepstral coefficients and the recovery of all possible regions in the following sections.

Comparison of natural and synthetic cepstral vectors

The objective is to enable the comparison of natural and synthetic cepstral vectors. This comparison cannot be carried out directly because the source is not taken into account in the synthetic spectra. Furthermore, the mismatch between the speaker's vocal tract and the articulatory model geometries, even after adaptation, introduces some deviation between the two cepstral vectors.

Traditionally, a lifter (Meyer *et al.*, 1991) is applied to attenuate the contribution of first coefficients linked to the spectral tilt and last coefficients linked to the harmonics of the excitation source. However, a closer examination of natural speech spectra and those produced by the articulatory synthesizer shows that spectral peaks are also slightly shifted in frequency. Since we have at our disposal X-ray films and corresponding audio signals for one speaker it is possible to compare synthetic and natural spectra corresponding to the same vocal tract shapes and to investigate two kinds of transformation:

(i) affine transformation of cepstral coefficients,

(ii) warping frequency via bilinear transformation.

Cepstral adaptation

The comparison between real and synthetic data has been studied by Mokhtari (Mokhtari *et al.*, 2004) in a 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 transformation, one for each formant frequency and bandwidth, were derived from a set of five vowels. Similarly, we are considering affine transformations to bring cepstral coefficients of natural and synthetic speech closer. 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 approximated by :

$$c'_n \approx a_n.c_n + b_n \tag{6}$$

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 :

$$E_n = \sum_k \left|\left|c'_{nk} - (a_n.c_{nk} + b_n)\right|\right|^2 \tag{7}$$

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

It turns out that the adaptation decreases the spectral distance between natural and synthetic spectra. However, this improvement is accompanied by a flattening of the spectrum, which thus reduces the emergence of formant peaks (see figure 1). The affine adaptation for the very first cepstral coefficients captures sthe spectral tilt without destroying the formantic

structure. Frequency warping described below is used to get a better fitting between spectral peaks.

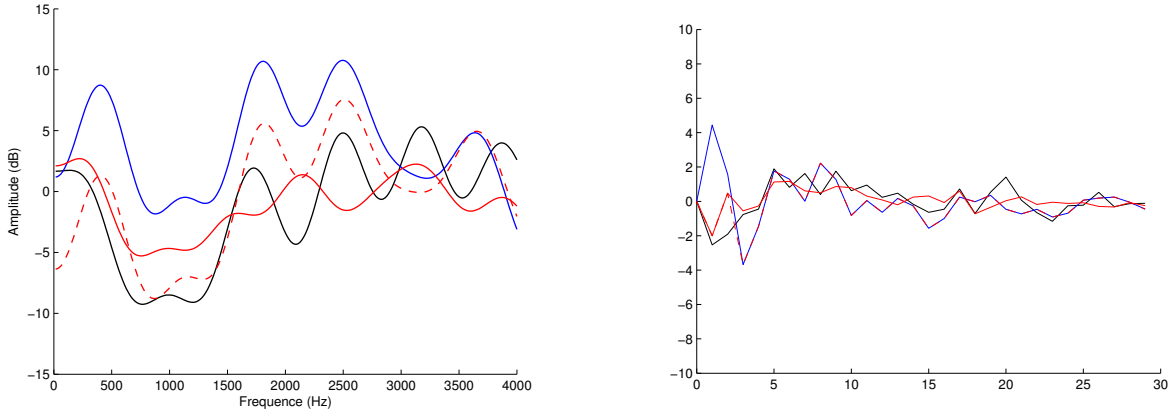

FIGURE 1: Cepstral coefficients ans cepstrally smoothed spectra before and after the adaptation with 30 ans two coefficients : natural (blue line), synthetic (black line), adaptation of 30 coefficients (red line) and 2 coefficients (red dashed line).

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 centerline used to decompose the vocal tract into elementary tubes. Indeed, the bilinear transform is a classical tool to perform vocal tract length normalization. It gives the new frequency variable z_{new} according to the following expression:

$$z_{new} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad -1 < \alpha < 1 \tag{8}$$

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

$$\omega_{new} = \omega + 2\tan^{-1}\frac{\alpha \sin\omega}{1 - \alpha\cos\omega} \tag{9}$$

The coefficient α has been adjusted so as to minimize the deviation between peaks of natural and synthetic spectra for two X-ray films (see figure 2).

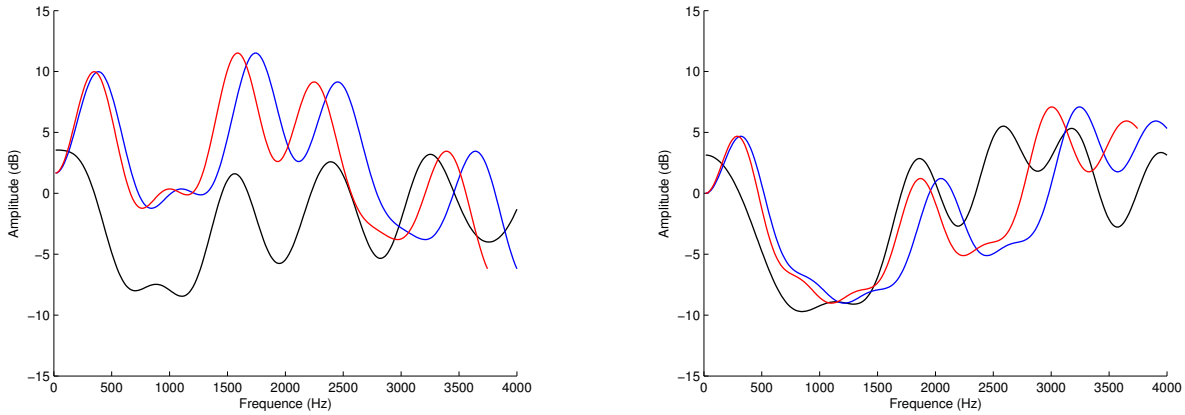

FIGURE 2: Cepstrally smoothed spectra : natural (blue line), synthetic (black line) and after warping (red line).

Inversion is all the faster so few cuboids of the codebook are explored. In fact, the quadratic programming method used to realize the inverse mapping in a cuboid is expensive in time. So cuboids without solutions must be discarded as early as possible.

Codebook organization

The idea is to use spectral peaks, i.e. mainly formants, to select relevant cuboids. However, spectral peaks are only used as indexing features and not as acoustic features to be optimized during inversion. The acoustic image of each articulatory cuboid is included in a parallelipiped in the space of the first three formants. This parallelipiped is the smallest parallelipiped containing the acoustic images of all the vertexes of the articulatory cuboid. Since the frequency domain of the second formant is the widest this formant is used as the primary indexing feature. This means that only articulatory cuboids with a F2 formant that fits a peak of the natural spectrum are considered.

Matching peaks of natural and synthetic spectra

After the selection based on F2, the objective is to find all the articulatory cuboids, which may correspond to an input cepstral vector. This search is carried out by matching peaks of the natural spectrum with formants of a synthetic spectrum. This matching must be robust against the presence of spurious peaks, or conversely the absence of some peaks, and to frequency shifts. This lax matching is realized via dynamic programming applied on spectral peaks.

Let $P = \big[p(m)\big] = p(1)\dots p(m)\dots p(M)$ and $Q = \big[q(n)\big] = q(1)\dots q(n)\dots q(N)$ be the sets of real and synthetic peaks, where $p(m)$ (resp. $q(n)$) corresponds to the m^{th} (resp. n^{th}) peak with M (resp. N) the number of peaks of P (resp. Q). The purpose is to extract subsets of P and Q represented by a sequence of indexes:

$$I = \big[i(k)\big] = i(1)\dots p(k)\dots p(K) \text{ with } K \leq M$$

$$J = \big[j(k)\big] = j(1)\dots j(k)\dots j(K) \text{ with } K \leq N$$

which maximizes the similarity of peaks.

I and J have to preserve the monotony of i and j: $i(k) < i(k+1)$ and $j(k) < j(k+1)$. The sequences of peaks $\bar{P}$ and $\bar{Q}$ correspond to the sequence of indexes I and J.

$$\bar{P} = \big[p(i(k))\big] = p(i(1))\dots p(i(k))\dots p(i(K))$$

$$\bar{Q} = \big[q(i(k))\big] = q(i(1))\dots q(i(k))\dots q(i(K))$$

The determination of i and j requires a local criterion which minimizes the distance between two peaks of P and Q. We chose the distance in frequency:

$$d\big(p(i(k)), q(j(k))\big) = \big|p(i(k)) - q(j(k))\big|$$

The overall criterion D is defined by:

$$D = \min_{K,I,J} \sum_{k=1}^{K} d\big(p(i(k)), q(j(k))\big) - B \tag{10}$$

where B is a positive bonus term chosen in order to favor the matching of strong peaks, i.e. formants. The bonus prevents the minimization from returning an empty matching between both spectra. This problem is solved by dynamic programming. Let $D(m,n)$ be the partial measure defined by:

$$D(m,n) = \min_{i,j} \sum_{k=1}^{k^*} d\big(p(i(k)), q(j(k))\big) - B \tag{11}$$

with $i(k^*) = m$ and $j(k^*) = n$. The sum is split into two parts:

$$D(m,n) = \min_{i,j} \left\{ d\big(p(i(k^*)), q(j(k^*))\big) - B + \sum_{k=1}^{k^*-1} d\big(p(i(k)), q(j(k))\big) - B \right\} \quad (12)$$

If $i(k^*-1) = l_1$ and $j(k^*-1) = l_2$, the recursive formula is given by:

$$D(m,n) = \min_{l_1<m, l_2<n} \left\{ d\big(p(m), q(n)\big) - B + D(l_1, l_2) \right\} \quad (13)$$

D is thus a pseudo distance, which enables a lax comparison between an unknown input spectrum and all the synthetic spectra computed at the cuboids centers. In addition to increasing the robustness against spurious peaks the lax comparison also allows the range of formant values corresponding to the acoustic image of a cuboid to be taken into account. Only one pseudo distance calculation for each cuboid is thus necessary.

It should be noted that the local solving (presented at section 2) via quadratic programming uses cepstral vectors whereas the search for possible cuboids uses cepstrally smoothed spectra.

EXPERIMENTS

The experiments exploit four X-ray films recorded by the same speaker. These four films were chosen because sound and image qualities are very good. The first two are a series of six short sentences ranging from /se dø si yltɛR/ to /se dø sikst skyltɛR/ (each sentence contains one more non-labial consonant between /i/ and /y/ than the previous one) at normal and fast speech rates. The last two are a series of /VCV/ /aku iku uku atu itu utu/ at normal and fast speech rates. Despite their very limited duration (37 s of speech in total) the articulatory variability turned out to be sufficient to construct an articulatory model which covers another speaker very precisely as shown in (Laprie and Busset, 2011). Compared to X-ray microbeam or EMA data, the whole 2D geometry of the vocal tract is known in X-ray images. This enables the geometric precision of inversion to be evaluated on the whole vocal tract, and especially the whole tongue contour, rather than only on 3 or 4 sensors glued on the front part of the tongue.

The linear articulatory model uses one parameter for the jaw opening, four for the tongue, one for the lips and one for the larynx (Busset and Laprie, 2011). The preliminary step of inversion consisted of building the codebook. The inversion presented in this paper is applied on each frame independently without taking the context into account. Indeed, the objective was not to reconstruct articulatory trajectories but only to test the recovery of vocal tract shapes from the speech signal. Only vowels were inverted since the acoustic simulation used here only covers vowel sounds with a sufficient quality. In total we thus inverted 133 speech signal windows. The pruning using F2 enables 32% of the codebook to be discarded. Then, after the lax peak matching only 5% of the cuboids of codebook remained on average. Finally, inversion via the quadratic programming method was applied on these cuboids.

We tested inversion of real data ; LPC spectra and linear cepstral coefficients were calculated on 32 ms Hamming windowed signals. The bilinear frequency warping was applied on LPC spectra. Next the cepstral adaptation using affine transformation is applied on the cepstral coefficients stemed from spectra after frequency warping. The transformation was made only on the two first coefficients because the use of more coefficients is accompanied by a flattening of the spectrum.

Before testing the inversion with an adaptation of the input acoustic vector, we performed inversion on cepstral vectors computed on spectra. In this case, the inversion gives no solution which was expected because of differences between natural and synthetic cepstral vectors. Table 1 represents the geometric and acoustic errors with or without cepstral adaptation. The adaptation is the affine transform given by the inversion of all the images corresponding to vowels. The acoustic error is defined as the euclidean distance between the input cepstral vector and the inverted one.

Without adaptation		With adaptation	
Geometrical error (mm)	Acoustic error	Geometrical error (mm)	Acoustic error
1.2	6.20	1.0	3.82

TABLE 1: Average errors for the inversion of the vowels in the corpus. At the left, inverse solutions are looked for without cepstral adaptation (only with frequency warping) and at the right with cepstral adaptation.

First of all, we observe that inversion is possible when the input vector is adapted. The affine transform of the first two cepstral coefficients after frequency warping improves the results. In fact the frequency warping moves the spectral peaks and the affine transform allows the adjustment of the spectral tilt. The figure 3 shows an example of inversion with the corresponding shape ; the inverted shape is the one which minimize the geometric distance.

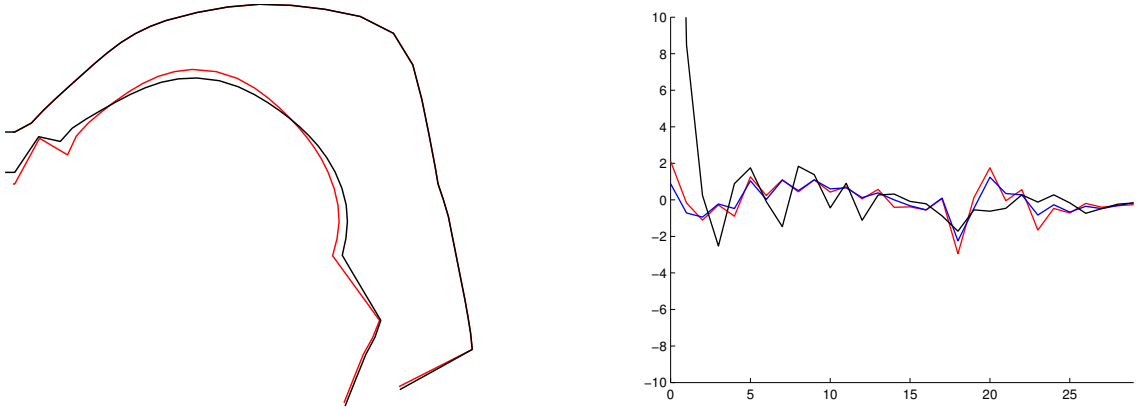

FIGURE 3: Vocal tract and cepstral coefficients for the inversion of one frame for the vowel /ɛ/. The black line corresponding to the vocal tract extracted from the image and the corresponding cepstral coefficients. The red line is the best solution (that one which minimize the geometric error) and the blue is the re-synthesis of the inverted articulatory parameters.

CONCLUSION

Our objective was to investigate the access to an articulatory codebook using cepstral coefficients in the favorable case where the geometrical mismatch between the articulatory model and the speaker's vocal tract is minimal. However, it should be stressed that the articulatory model has been kept voluntarily sufficiently rough so as not to capture speaker specific anatomical details.

These experiments validate the access to an articulatory codebook using cepstral coefficients. In particular, the geometric evaluation covers the whole tongue and not only the front part of the tongue where EMA or microbeam sensors are usually glued. This means that no artificial compensation can be used to force the fitting in the front part of the tongue to the detriment of the realism of the whole tongue shape especially in the tongue root region as it sometimes happens in other inversion methods (Panchapagesan and Alwan, 2011).

This work was dedicated to the key step of inversion which consists of recovering vocal tract shapes that may have been used to generate a speech sound. The future work will focus the reconstruction of articulatory trajectories. The second aspect that will be studied is the influence of the model mismatch about inversion results by artificially degrading the goodness of fit between the model and the speaker's vocal tract geometry.

REFERENCES

Busset, J. and Laprie, Y. (**2011**). "Adaptation of cepstral coefficients for acoustic-to-articulatory inversion", in *The Ninth International Seminar on Speech Production - ISSP'11* (Canada, Montreal).

Goldfarb, D. and Idnani, A. (**1983**). "A numerically stable dual method for solving strictly convex quadratic programs", Mathematical Programming **27**, 1–33.

Laprie, Y. and Busset, J. (**2011**). "Construction and evaluation of an articulatory model of the vocal tract", in *19th European Signal Processing Conference - EUSIPCO-2011* (Barcelona, Spain).

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

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

Narayanan, S., Bresch, E., Ghosh, P. K., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., and Zhu, Y. (**2011**). "A multimodal real-time mri articulatory corpus for speech research", in *12th Annual Conference of the International Speech Communication Association - INTERSPEECH 2011* (Florence).

Oppenheim, A. and Johnson, D. (**1972**). "Discrete representation of signals", Proceedings of the IEEE **60**, 681 – 691.

Ouni, S. and Laprie, Y. (**2005**). "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion", JASA **118**, 444–460.

Panchapagesan, S. and Alwan, A. (**2011**). "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model", The Journal of the Acoustical Society of America **129**, 2144–2162, URL `http://link.aip.org/link/?JAS/129/2144/1`.

Potard, B. and Laprie, Y. (**2007**). "Compact representations of the articulatory-to-acoustic mapping", in *Interspeech, Antwerp*.

Potard, B. and Laprie, Y. (**2009**). "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).

Sock, R., Hirsch, F., Laprie, Y., Perrier, P., Vaxelaire, B., Brock, G., Bouarourou, F., Fauth, C., Hecker, V., Ma, L., Busset, J., and Sturm, J. (**2011**). "DOCVACIM an X-ray database and tools for the study of coarticulation, inversion and evaluation of physical models", in *The Ninth International Seminar on Speech Production - ISSP'11* (Canada, Montreal).