Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

Technical Memo
07-02

A review of state-of-the-art speech modelling methods for the parameterisation of expressive synthetic speech

Sacha Krstulović

March 2007

Deutsches Forschungszentrum für Künstliche Intelligenz

Postfach 20 80
67608 Kaiserslautern, FRG
Tel.: + 49 (631) 205-3211
Fax: + 49 (631) 205-3210
E-Mail: info@dfki.uni-kl.de

Stuhlsatzenhausweg 3
66123 Saarbrücken, FRG
Tel.: + 49 (681) 302-5252
Fax: + 49 (681) 302-5341
E-Mail: info@dfki.de

WWW: http://www.dfki.de

Deutsches Forschungszentrum für Künstliche Intelligenz

DFKI GmbH

German Research Center for Artificial Intelligence

Founded in 1988, DFKI today is one of the largest nonprofit contract research institutes in the field of innovative software technology based on Artificial Intelligence (AI) methods. DFKI is focusing on the complete cycle of innovation — from world-class basic research and technology development through leading-edge demonstrators and prototypes to product functions and commercialization.

Based in Kaiserslautern and Saarbrücken, the German Research Center for Artificial Intelligence ranks among the important “Centers of Excellence” worldwide.

An important element of DFKI’s mission is to move innovathons as quickly as possible from the lab into the marketplace. Only by maintaining research projects at the forefront of science can DFKI have the strength to meet its technology transfer goals.

The key directors of DFKI are Prof. Wolfgang Wahlster (CEO) and Dr. Walter Olthoff (CFO).

DFKI’s six research departments are directed by internationally recognized research scientists:

- Image Understanding and Pattern Recognition (Director: Prof. Thomas Breuel)
- Knowledge Management (Director: Prof. A. Dengel)
- Intelligent Visualization and Simulation Systems (Director: Prof. H. Hagen)
- Deduction and Multiagent Systems (Director: Prof. J. Siekmann)
- Language Technology (Director: Prof. H. Uszkoreit)
- Intelligent User Interfaces (Director: Prof. W. Wahlster)

Furthermore, since 2002 the Institute for Information Systems (IWi) (Director: Prof. August-Wilhelm Scheer) is part of the DFKI.

In this series, DFKI publishes research reports, technical memos, documents (eg. workshop proceedings), and final project reports. The aim is to make new results, ideas, and software available as quickly as possible.

Prof. Wolfgang Wahlster
Director

A review of state-of-the-art speech modelling methods for the parameterisation of expressive synthetic speech

Sacha Krstulović

DFKI-07-02

This work has been supported by a grant from the Deutsche Forschungsgemeinschaft for project PAVOQUE.

ISSN 0946-0071

A review of state-of-the-art speech modelling methods for the parameterisation of expressive synthetic speech

Sacha Krstulović

March 2, 2007

Contents

1 General and introductory considerations

This document will review a sample of available voice modelling and transformation techniques, in view of an application in expressive unit-selection based speech synthesis in the framework of the PAVOQUE project. The underlying idea is to introduce some parametric modification capabilities at the level of the synthesis system, in order to compensate for the sparsity and rigidity, in terms of available emotional speaking styles, of the databases used to define speech synthesis voices.
For this work, emotion-related parametric modifications will be restricted to the domains of voice quality and prosody, as suggested by several reviews addressing the vocal correlates of emotions (Schröder, 2001; Schröder, 2004; Roehling et al., 2006).
The present report will start with a review of some techniques related to voice quality modelling and modification. First, it will explore the techniques related to glottal flow modelling. Then, it will review the domain of cross-speaker voice transformations, in view of a transposition to the domain of cross-emotion voice transformations. This topic will be exposed from the perspective of the parametric spectral modelling of speech and then from the perspective of available spectral transformation techniques. Then, the domain of prosodic parameterisation and modification will be reviewed.

2 Parameterisation of voice quality

2.1 Techniques based on explicit glottal flow models

The characteristics of the glottal source are known to play an important role in the definition of voice quality (Laver, 1980). As a consequence, it is natural to try and define the parameterisation and modification of voice quality in relation to measurements of the glottal flow from the speech wave.

Existing methods – In this area, significant works include, in chronological order:

- (Liu et al., 1992): the Lin-Fant glottal flow model (Fant et al., 1985) is iteratively fitted to the residual of an adapted inverse filtering model. Two versions of the algorithm are proposed: one where the vocal tract inverse filter is estimated by pitch-synchronous LPC analysis and the parameters of the glottal flow model are fitted to the residual by a standard quadratic programming procedure, and one where the vocal tract inverse filter and the glottal source parameters are jointly optimised by an iterative analysis-by-synthesis method. The quality of this method is measured in terms of signal reconstruction error.

- (Alku, 1992): the Pitch-Synchronous Iterative Adaptive Inverse Filtering (PSIAIF) method designed in this work uses nested LPC-based filtering stages, with different orders, to inverse-filter the glottal wave effects independently of the vocal tract effects. This method seems successful in recovering the artifical excitation of some synthetic vowels, and delivers some "realistic" glottal flow shapes on voiced natural vowels, but neither a comparison with human glottal flows (e.g., laryngograph signals) nor a study of the method's results on unvoiced speech are available. Moreover, the method has the drawback that for certain phonation types, such as pressed or breathy, the glottal filter locks on the first formant instead of the glottal formant. However, a practical advantage of this method is that a Matlab-based open-source implementation is available (Airas et al., 2005).

- (Akande and Murphy, 2005): this work aims at alleviating the glottal formant estimation problem observed by (Alku, 1992), by using the phase as a discriminant information between the glottal formant and the first formant. Again, it uses a series of filtering operations: a stage of low-pass filtering, aiming at the partial separation of the effect of glottal flow from speech, is followed by the estimation of a vocal-tract filter, which is constrained in terms of unit gain at the first formant, minimal phase delay for the whole filter and bandwidth to frequency ratios for the available formants. These constraints are achieved by adjusting two parameters for the filter estimation, namely the window length and the filter order. The inverse filtering of speech by the estimated vocal tract filters shows some "realistic" glottal flow shapes for voiced speech.

The above methods aim at recovering a time-domain signal which describes the glottal flow. However, in view of voice quality modification, there is a need to fit a *control model* over the glottal wave shapes to relate them to voice quality or speaker categories, in view of modification or classification. Significant works in this direction include:

- (Childers, 1995): some parametric glottal flow models (the polynomial model from (Milenkovic, 1993) or the Lin-Fant model (Fant et al., 1985)) are fitted to some Linear Prediction residuals, using a cross-correlation based method. Then, the estimated glottal waveforms are quantised and reduced to codebooks for voiced speech. The voiced codebooks are complemented by unvoiced codebooks estimated according to the usual CELP coding techniques (Schroeder and Atal, 1985; Schultheiss and Lacroix, 1989), i.e., without resorting to a glottal wave model. Voice quality modification is performed by switching the codebooks resulting from the separate analysis of a speaker's voice at various voice quality regimes, e.g., switching the breathy voice excitation codebook with the modal voice excitation codebook. Formal listening tests are used to validate the approach.

- (d'Alessandro and Doval, 1998): an estimation of the glottal wave based on the above-mentioned PSIAIF method (Alku, 1992) is followed by a periodic/aperiodic decomposition of the source signal and the estimation of various parameters of the resulting analytic model of the glottal source spectrum. These parameters are: open quotient, glottal formant, spectral tilt, phase and aperiodic component. A set of rules is defined over these parameters to provide voice quality modifications. A limited number of audio examples, illustrating some voice quality modifications performed in the framework of a MBROLA-based TTS synthesis system, are provided to the reader, as an informal listening test. According to a personal communication from the author, no open-source implementation of the method is yet available.

- The above seems to have been work in progress, since (d'Alessandro and Doval, 2003) specify some alternative glottal spectrum control parameters for voice quality modification, namely: in the time domain, amplitude of voicing, fundamental frequency, open quotient, asymmetry coefficient, return phase coefficient; in the spectral domain, fundamental frequency, frequency locations of spectral peaks, amplitude of spectral peak, quality coefficient of the spectral peak, spectral tilt. Although the article indicates that high-quality voice modifications can be achieved, no formal listening tests are performed, and the article concludes on the difficulty of finding a set of rules for performing meaningful modifications.

- (Mokhtari et al., 2003): the glottal flow waveform is determined by estimating and removing the formant contribution from the speech spectrum. The formants are

estimated using a method based on the Linear Prediction Cepstrum Coefficients (LPCCs). The estimation of the glottal flow shape is performed for a single glottal flow period, by inverse formant-filtering and a series of other filtering operations and rules, and is restricted to some "centres of reliability", which are automatically located with respect to some syllabic structure considerations. A Principal Components Analysis (PCA) is performed on the obtained sets of individual glottal wave shapes to find voice quality correlates expressed a linear combinations of shape factors. The method is evaluated in the framework of a voice quality classification task based on dimensionality reduction and decision trees, using tape-recorded natural speech data issued from a reference study of voice quality (Laver, 1980). The sparsity and tape-recording quality of the test data make it difficult to assess the validity of the exposed results.

- (Lugger et al., 2006): a heavily stylised model of the glottal spectrum is estimated according to the following method. A pitch-tracking is performed using the RAPT algorithm (Talkin et al., 1995) and the formant locations are estimated according to the LPC-based algorithm designed by (Talkin, 1987), both implemented as part of the ESPS software toolkit. The influences of the formants on the spectrum are estimated according to Fant's model of the speech acoustics (Fant, 1970), and then subtracted from the speech spectrum. The description of the resulting residual spectrum is then reduced to 5 gradient values computed at the locations of the fundamental and formant frequencies. The method is evaluated in the framework of several LDA-based classification tasks, dealing with gender, phonation types (voice qualities) or emotions, with apparently successful results under real-world recording conditions. The same method is reported to have given good results for the automatic discrimination of pathological voices in some anterior work (Wokurek and Pützer, 2003).

Conclusion – From the above examples, glottal flow modelling and its application to the characterisation of voice quality seems to remain an open field of research. There does not appear to be a standard practice or a "best model", especially regarding the voice quality parameters that should be used to control the glottal flow model. For example, in (Alku, 2003), a variety of glottal flow parameterisation methods are reviewed, but the author concludes that:

> "There is hardly any single method that would outperform all the others".

He then advocates the use of the Open Quotient (OQ) or the Normalised Amplitude Quotient (NAQ) if one would like to reduce the description of the glottal wave to a single coefficient, while remarking that:

> "once the original time-domain pulseforms are expressed using any parameterisation method, which is in the form of a single numerical value, a major part of the original information embedded in the waveforms will be thrown away".

Globally, most of the glottal flow extraction studies are restricted to the analysis of voiced speech and sustained vowels, but the behaviour of, e.g., OQ or NAQ measurements in the unvoiced parts of natural speech is not well identified. Finally, apart from the notable exception of the Matlab-based IAIF software (Airas et al., 2005), there does not seem to be a lot of open-source software available to test the existing glottal flow models.
Nevertheless, the PAVOQUE database may be of interest to researchers in this area, in order to help with the assessment of the existing models, or to learn some voice quality-related glottal flow control parameters from real speech data.

Given the difficulties attached to the use of explicit glottal waveform models, it appears preferable to make only minimal modelling assumptions as far as the glottal phenomena are concerned. For instance, one can consider that a transformation of voice quality or glottal characteristics corresponds to a global transformation of the speech spectrum, operated in an adequate spectral domain which will deal with the glottal flow modelling aspects in a more explicit way. Such a rationale is currently widely applied in the domain of cross-speaker voice transformation.

2.2 Voice transformation as a global acoustic transformation

Most of the existing cross-speaker voice transformation techniques are based on global spectral transformations, and could easily be transposed to emotion-related voice quality transformation. The use of global spectral transforms for cross-speaker voice quality conversion has been supported by many works, such as the ones reviewed in (Kuwabara and Sagisaka, 1995), where the authors conclude that:

> "As far as speaker individuality is concerned, there is no single acoustic parameter that carries the entire information. Dominant acoustic features depend both on the speaker and on the speech material to be examined. This leads us to conclude that whereas it may not be feasible to take a simple parametric approach to changing voice individuality, developments in speech technology have made it possible to change the individuality from one speaker to another without explicit modelling of a speaker's voice characteristics by using acoustic features directly. This technology seems to be promising as far as speaker conversion is concerned, though it still leaves much room for improvement with respect to the quality of the converted speech and the manipulation of its prosody."

In this line of idea, notable works in the domain of cross-speaker voice transformation include:

- (Valbret et al., 1992), who implement voice transformations as a mix of prosodic transformations, based on the PSOLA technique, and spectral transformations, inspired from the seminal work of (Abe et al., 1988). The spectral transformations are implemented as a set of linear transforms, operated in a parametric cepstral domain (LPCCs), between the source and the target speaker's spectra. The set of transforms is determined by a partition of the speaker's spectral spaces, obtained by Vector-Quantisation.

- In a related manner, (Stylianou et al., 1998) implement Voice Transformation as a linear transformation, operated in a different parametric Mel-cepstral domain, which maps the two speaker's spectral subspaces in a more global way. The speaker's subspaces are modelled by global Gaussian Mixture Models (GMMs) instead of locally isolated partitions. In this particular work, no PSOLA-based prosodic modification is performed, in contrast to an anterior work by the same authors (Stylianou et al., 1995).

- In (Kain and Macon, 1998), a similar method is employed, but in the Line Spectral Frequency domain, and through the estimation of a single GMM for the two speakers. In a follow up work (Kain and Macon, 2001), a transformation of the LPC *residual* is added in the form of a prediction from the spectral shapes, with an increase in the resulting synthetic speech quality. An overall degradation of the quality with respect to real speech is nevertheless noticed after the voice transformation.

- (Arslan, 1999) proposes a voice conversion algorithm based on "Segmental Codebooks" formed in the Line Spectrum Frequency (LSF) domain. The idea is to quantise the source and target speaker's subspaces by aligning HMMs to the speech data, and then considering the mean of each state as a codeword. Unseen speech frames are coded as a linear combination of codewords, and voice transformation is performed by preserving the weights vector while switching the codebook to that of the target speaker. A mapping of the filter excitation is also provided in the frequency domain, and the algorithm is completed by additional layers of modification for the of bandwidth, the pitch-scale, the duration-scale and the energy-scale. The results are evaluated by listening assessment but also, notably, by application of some automatic speaker recognition techniques.

  In a follow-up work, (Turk and Arslan, 2006) identify several factors that may hamper the performances of the algorithm, particularly when the spectral shapes to be matched between the source and the target codebook are too far from each other. Therefore, they propose some methods to prune such ill-defined cases out of the codebooks. Both works introduce several interesting ideas, but the published sound examples sound quite far from natural. It is difficult to know if this method sounds better or worse than any of the anterior ones, since no audio examples are available for the older methods.

- A number of voice adaptation techniques have been proposed in the framework of HMM-based speech synthesis (Yoshimura et al., 1999; Masuko et al., 1997; Yoshimura et al., 1997; Tamura et al., 1998; Shichiri et al., 2002). They are affiliated with speaker adaptation for speech recognition, and will be reviewed in the section 2.4 of the present document.

The above works have been applied to voice transformation between speakers, not between voice qualities within the same speaker. A notable exception is the work of (Turk et al., 2005), where spectral shapes related to voice quality have been linearly interpolated in the LSF domain to produce intermediate voice quality levels, with good results in terms of subjective perception. *This result supports the generalisation of the cross-speaker voice transformation techniques to within-speaker, emotion-related modifications of the voice quality.*

However, the acoustic changes related to voice quality variations could be expected to have a different nature in the distinct cases of cross-speaker transformations and within-speaker/cross-emotion transformations. On can surmise that cross-speaker transformations encompass single-speaker/cross emotion transformations because they are designed to model the speaker variability across a whole range of speaking styles. However, there is no guarantee that the reported speaker transformation methods have been tested on anything else than neutral speech, both on the source and on the target speaker sides, and without an explicit and wide sampling of (possibly emotional) speaking styles. In our own experiments, cross-speaker variability will be eliminated, but a wider sampling in terms of emotions and expressiveness will possibly bring its own lot of specific acoustic modelling and mapping problems.

From a more global point of view, the review of these voice transformation approaches suggests that two choices are crucial for the definition of the transformation method: the choice of the spectral feature space and the choice of the mapping method. These two aspects will be developed in the following sections.

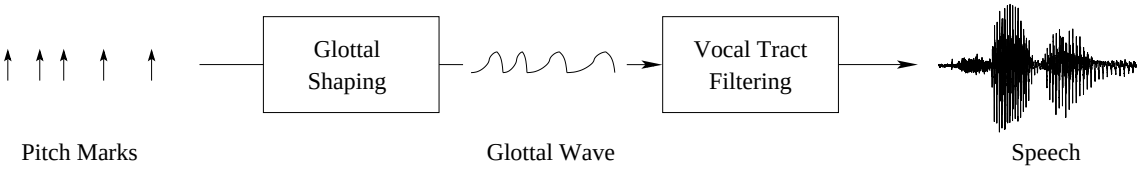

Figure 1: A two-stage speech production model.

2.3 Speech analysis and choice of the acoustic feature space

Whereas the common point in the above-mentioned techniques is to establish linear transformations between the acoustic spaces of two speakers, a notable variation is the choice of the acoustic feature space, which can be either Cepstral Coefficients or Line Spectrum Frequencies (LSFs).

2.3.1 LPCCs versus LSFs

Linearisation advantage of Cepstral Coefficients – The main advantage of the Cepstral Coefficients is found in their linearisation properties. If we suppose that the production model underlying speech is an impulse train source, locating the glottal closure instants, fed to a "glottal shaping" filter to produce a glottal wave, which will in turn be fed to a "vocal tract" filter in charge of imposing the formant structure (figure 1), then the effects of the glottal filter and of the VT filter are convolved in time. Hence, they are multiplied in the frequency domain, and therefore they are additive in the log-domain. It has been indeed long known (Oppenheim and Schafer, 1968; Schafer and Rabiner, 1990) that such convolutional effects can be analysed as additive effects in the cepstral domain. The success of Cepstral Mean Subtraction (Mokbel et al., 1994) or speaker adaptation techniques (Woodland, 2001), usually applied in the non-parametric cepstral domain (MFCCs) for speech recognition experiments, illustrates the validity of this consideration. Assuming that the variabilities related to the glottal source on the one hand, and to the vocal tract configuration on the other hand, are linearly independent in the cepstral domain is an important property if ones wishes produce a linear analysis of the variability along the distinct axes of the vocal tract effects and the glottal flow effects.

Stability and interpolation advantages of LSFs – However, in an experiment produced for the needs of speech coding, (Paliwal, 1995) remarked that the interpolation of Linear Prediction Cepstral Coefficients (LPCCs) can produce unstable Auto-Regressive filters in the equivalent LPC spectral domain, in around 5-7% of the frames for a fixed-frame 10-20ms speech analysis rate corresponding to the human pitch range. He thus advocates the use of LSFs, for which interpolation produces approximately the same level of distortion than with LPCCs, but no unstable filters by definition. (Kain and Macon, 1998) add arguments in favour of LSFs, by remarking that the effect of varying each individual LSF value is localised in frequency, for a better robustness to interpolation errors, and that the LSFs relate well to the formant location and bandwidths, which are perceptually relevant for voice characteristics.

Open questions – Yet, there is no widespread consensus as to whether transformations between voice subspaces in the frequency domain (or LSF-based formant domain) are linear or non-linear in nature, whereas linear transformations in the cepstral domain can be related to convolutional effects or perceptual distances. Indeed, LSFs have not been much

used in the framework of speech or speaker recognition, but that is probably because they don't easily lend themselves to the compensation of the convolutional channel effects, the latter playing a major role in error reduction for speech or speaker recognition (Reynolds, 1994). In addition, the study of (Paliwal, 1995) was conducted at a fixed frame rate for undifferentiated speech, and it would be interesting to investigate if the unstable LPC filters could still be observed when using a pitch-synchronous analysis, and if yes, if their appearance would be limited to certain phonetic classes or would be related to a particular acoustic phenomenon. A post-processing method could also be used to correct the unstable frames in the LPCC interpolation case, e.g., by thresholding the values of the equivalent reflection coefficients, or by relocating the equivalent poles inside the unit circle. *As a conclusion, it would be desirable to compare the performances of both LSFs and LPCCs for the need of our own experiments.*

2.3.2 A note on high-quality LPC modelling related to spectral envelope models

The classical LPC estimation techniques, such as the Autocorrelation method or the Itakura-Saito method, are known to suffer from a bias in the localisation of the LPC poles for high-pitched voiced speech, e.g., for female voices or singing voices: the poles tend to lock on the speech harmonics rather than on the formants. This non-linear effect is undesirable for several reasons:

- for high-quality voice modification, it may be desirable to relocate the speech harmonics or to modify the nature of the vocal excitation while preserving the shape of the spectral envelope which is "truly" related to the vocal tract resonances;
- the biased spectral envelope estimates can potentially introduce errors when trying to establish a mapping between two voice subspaces, because erroneous estimates could appear on a single side of otherwise related zones of both subspaces, e.g., when trying to transform a high-pitched voice into a low-pitched voice through a phoneme-dependent mapping.

As a result, several methods have been designed to achieve better spectral envelope modelling while still relying on a Linear Prediction model. The underlying principle generally consists in fitting a LPC envelope over a finite set of spectral peaks (Jaroudi and Makhoul, 1991; Galas and Rodet, 1991; Cappé et al., 1995; Villavicencio et al., 2006), according to various criteria. The spectral peaks can come either from the peak-picking of the harmonics in a FFT (Valbret et al., 1992; Villavicencio et al., 2006), or from a full-fledged sinusoidal analysis (Macon, 1996; Stylianou et al., 1998), or from any kind of manual or automatic specification (Galas and Rodet, 1991; Cappé et al., 1995). Alternately, the spectral envelope can be fitted directly from a set of non-parametric Mel-Filterbank Cepstral Coefficients (Imai, 1983). These methods claim to provide better spectral envelope models, although at the cost of an increased theoretical and computational complexity. They have spread in the domain of high-quality voice transformation due to their better envelope-modelling performances, but to our knowledge they have not been applied into the domain of speech or speaker recognition.
Indeed, the standard practice in the domain of speech/speaker recognition is to make direct use of Mel-Filterbank Cepstrum Coefficients (MFCCs) instead of LPCCs, mostly because MFCCs are based on the FFT and hence are less costly to compute than a LPC estimation. In addition, the use of the Mel-scale is justified by some perceptual considerations, though its effect has been proved negligible as far as speaker recognition scores are concerned (Reynolds, 1994). But, it is impossible to recover the original FFT-based spectrum from

the MFCCs because the filterbanks operate a non-invertible integration of the spectral samples. In contrast, a constraint that we face in speech synthesis is the necessity to re-synthesise the analysed and transformed speech samples. If one wants to re-synthesise a signal analysed in the cepstral domain, he is bound either to use some LPC-derived cepstral coefficients (LPCCs) from the beginning, or to resort to a LPC model derived from one of the above-mentioned envelope-fitting methods.
As a matter of fact, the Mel Log Spectrum Approximation (MLSA) filters (Imai, 1983) define a LPC-based spectral envelope from the Mel cepstrum. They have been originally used to synthesise speech with artifical excitations, and are still used in this way for HMM-based synthesis (Yoshimura et al., 1999). However, it is well-known that using synthetic excitations produces worse results than using excitations obtained from the inverse filtering of natural speech (Macchi et al., 1993), since the natural residuals retain and compensate for the speech-related information which escapes the LPC model. A workaround could therefore consist in using the MLSA filters as a LPC estimator rather than as a complete vocoding technique, by employing the resulting LPC filter as an inverse filter which could preserve a link to a complementary natural excitation.
On the practical side, as far as LPC estimation is concerned, the standard open source feature extraction packages (Edinburgh Speech Tools, HTK, ESPS, Spro) implement the autocorrelation method for LPC analysis, and they don't provide sinusoidal analysis modules nor envelope fitting modules. However, certain classes of sinusoidal analysis algorithms are equivalent to performing some Matching Pursuit (Macon and Clements, 1999) on the speech signal. As a matter of fact, the Matching Pursuit Toolkit (Krstulović and Gribonval, 2006) implements a fast version of this type of method. Fitting LPC envelopes to the sinusoidal estimates would correspond to adding an implementation of the method described in (Jaroudi and Makhoul, 1991), (Cappé et al., 1995) or (Villavicencio et al., 2006) to the package, and would represent a tractable amount of programming effort. *As a summary, envelope-based LPC estimation techniques would probably correspond to better "vocal filter" models, which may prove essential to the good performance of voice quality related spectral mapping, but at the expense of an increased computational complexity and programming effort.*

In addition to the choice of the feature space where the voice transformation is to be applied, the other choice of importance concerns the nature of the mapping between the source and the target voice characteristics.

2.4 Mapping of the spectral shapes for voice adaptation

2.4.1 General considerations

(Baudoin and Stylianou, 1996) have systematically compared several spectral mapping techniques, based on: Vector-Quantisation Codebooks, Gaussian-Mixture Models (GMM), Neural Networks and Linear Multivariate Regression. In all cases, the mapping is performed in the space of the Cepstral Coefficients resulting from the regularised estimation of the spectral envelope from (Cappé et al., 1995), applied to the harmonic part of the HNM model (Stylianou et al., 1995). Objective tests, in terms of normalised spectral distance from the source speaker (further is better) and the target speaker (closer is better) indicate that the methods performing a global mapping over the whole space, namely GMMs and VQ with weighted map, perform better than the methods which perform a class-dependent mapping or a one-to-one mapping in isolated parts of the vocal space. Subjective listening tests indicate that the GMM-based mapping gives the best results, although speech quality after transformation is globally judged very poor in all cases. Nevertheless, *these results suggest that voice variability is better modelled as a transformation between global*

models that cover the whole acoustic feature space for each speaker, rather than in terms of a template lookup applied over segregated zones of both speaker spaces.

2.4.2 A parallel with the adaptation and state tying techniques used in speech recognition

The above result can be parallelled with the success obtained with speaker adaptation methods in the HMM framework, both in speech recognition (Woodland, 2001) and in speech synthesis (Yoshimura et al., 1999; Masuko et al., 1997; Yoshimura et al., 1997; Tamura et al., 1998; Shichiri et al., 2002). As a matter of fact, Gaussian Mixture Models (GMMs) can be interpreted as a "soft" and global partition of the acoustic space in terms of a set of Gaussian classes, deployed in a multi-dimensional feature space. In this model, a single Gaussian represents the repartition of a data class in terms of second order statistics, i.e., by specification of a mean and a variance. The degree of belonging of a data point to a Gaussian class can be evaluated by computing a likelihood. For a mixture of Gaussians, every data point belongs to every class within a certain proportion of likelihood. Models made of sets of class-dependent GMMs are a way to introduce some supervision in the training process, while HMMs are a way to introduce some constraints on the sequential ordering of a set of GMMs. To sum up, GMMs and HMMs are based on a statistical modelling paradigm which amounts to a "soft" and global clustering of the acoustical space, and where the use of statistics allows every data point to be informative about every cluster. *GMMs and HMMs are therefore particularly well suited to the definition of global transforms between acoustic feature subspaces.*

Speaker Adaptation techniques – A review of the main speaker adaptation techniques known to date in the framework of HMM-based speech recognition can be found in (Woodland, 2001). The general idea behind these methods is to compensate the sparsity of speaker-specific data by using a large amount of speaker-independent data to build a generic speech model, and then to use the available amount of speaker-specific data, which is often small, to deduce a speaker-adapted model from the generic one. This consideration is to be parallelled with the necessary sparsity of the emotion or expression-specific data in the emotion-oriented synthesis databases.
The adaptation techniques currently known as state-of-the-art can be divided into 3 broad classes:

- *Maximum A-Posteriori (MAP) adaptation*: the parameters of the speaker-dependent model are obtained from the adaptation of the generic speech model through re-estimation formulae which include the generic model as a prior and which use the available speaker-specific data as new training samples. This method has the important property that it converges to the Maximum-Likelihood solution when the amount of adaptation data increases. In this framework, adaptation according to new data can re-organise the topology of the partition of the acoustic space, according to the amount of data that each Gaussian will "see";

- *Maximum-Likelihood Linear Regression (MLLR)*: the parameters of the speaker-dependent model are obtained through an affine transform applied to the parameters of the generic model. The parameters of the affine transform itself are estimated in a Maximum-Likelihood sense. In this framework, the generic model is adapted globally; in other terms, the Gaussians which don't "see" any data still undergo the affine transform, and the global arrangement of the Gaussians observed in the generic model is preserved after the transformation. This class of methods therefore gives

better results for sparser amounts of adaptation data, but it is usually outperformed by MAP adaptation when the amount of adaptation data increases;

- *Model Interpolation and Eigenvoices*: here, the generic model corresponds to a set of speaker-specific models, sometimes referred to as "anchor models", rather than a unique generic speech model. A new model for an unseen speaker is computed as a linear combination of the parameters of the anchor models, again according to a Maximum Likelihood paradigm with respect to a limited amount of adaptation data. This linear combination can be interpreted as an interpolation of the anchor models. In order to reduce the dimensionality of the linear transform, Principal Components Analysis (PCA) can be applied to the parameters of the anchor models, therefore reducing the set of anchor models to a smaller set of so-called eigenvoices. The eigenvoices double as an interesting analysis tool, since the PCA models the principal axes of variability[1] across the anchor speakers.

All these methods, in their respective way, operate a mapping between the model of a generic, speaker-independent acoustic space, and the model of a speaker-dependent acoustic subspace. Mapping a generic single-speaker model to a range of emotion-dependent acoustic subspaces could be understood as a similar problem, although the degree to which the above-mentioned methods can fit this interpretation remains subjected to a confrontation with the actual emotional data.

Tree-based state tying – Whereas the above-mentioned adaptation techniques aim at tackling the sparsity of data related to the speaker variability, another data sparsity problem arises when the variability in terms of phonetic contexts is considered. In the framework of speech recognition, this problem is usually tackled using tree-based state clustering and parameter tying methods (Young et al., 1994). These methods reduce the total number of model parameters by using decision trees to form clusters of Gaussian models, and by tying the Gaussian parameters on the basis of *a)* a set of questions related to some context-description features and *b)* a purity criterion defined as a likelihood measurement. This method presents an obvious parallel with the tree-based clustering operated in the pre-selection stage of unit-selection speech synthesis (Black and Taylor, 1997): the likelihood between the GMMs and the data is analogous to the acoustic distances between the units, and whereas the questions used in speech recognition deal mainly with immediate phonetic context in terms of left and right phoneme identities, speech synthesis uses a richer set of questions involving features such as accentuation, syllable related features etc. Although state-tying with a limited phonetic context has already been used for "trainable" speech synthesis (Donovan and Eide, 1998), the use of the whole range of questions applied in the framework of unit selection does not seem to have been extensively studied before, except in the framework of HMM-based synthesis (Tachibana et al., 2004). As a matter of fact, using the richer set of questions could prove useful in helping to achieve a better integration of the alignment and selection processes, thus leading to an overall better quality of the synthesis system, as compared to a plain phone-based alignment system. In addition, *emotional labels could be added to the context clustering procedure.* This possibility will be illustrated in the next section.

[1]PCA operates under a restrictive assumption of orthogonality between the factors of variability. Independent Components Analysis (ICA), which assumes only statistical independence between the factors, may lead to more meaningful results as far as the axes of variability are concerned. However, to our knowledge, ICA has never been applied to the analysis of speech variability.

2.4.3 Application of adaptation and context clustering into HMM-based speech synthesis

Speaker identity modification – Interestingly enough, the state-of-the-art speaker adaptation techniques have been systematically explored for voice adaptation in the framework of HMM-based speech synthesis:

- MAP adaptation is applied in (Masuko et al., 1997), with encouraging results assessed through listening experiments. This work underlines the need for a proper amount of adaptation data and a proper degree of state clustering;
- a Speaker Interpolation method, similar to the anchor models concept, is applied in (Yoshimura et al., 1997), showing encouraging perceptual results;
- MLLR for HMM-based synthesis is introduced in (Tamura et al., 1998) and is shown to provide a transformation quality comparable to MAP adaptation but with a much simpler setup, involving less control parameters. It is further developed in (Tamura et al., 2001) to include pitch and duration modelling in the adaptation process, in addition to the single adaptation of the spectral envelope.
- a preliminary investigation of the use of Eigenvoices is exposed in (Shichiri et al., 2002). In this work, some eigenvoices are computed from 10 speaker models, and the two first principal components are found to relate to the gender and sound volume characteristics of the synthetic voices. This suggests that eigenvoices represent a valid control model; however, no method is given to determine some adaptation parameters with respect to a new speaker's data.

These results seem all promising, but no systematic comparison of the performance of these adaptation schemes is available as far as speech synthesis quality or voice adaptation quality are concerned.

Speaking style modification – The above-mentioned ideas can be directly transposed to the domain of speaking style or expressiveness modification by considering that emotion-related voice alterations are equivalent to factors of voice identity, or to factors of phonetic context. As a matter of fact, adaptation techniques have recently been investigated in the domain of speaking style modification instead of voice identity modification in the framework of HMM-based synthesis:

- (Yamagishi et al., 2003) have investigated the modelling of four speaking styles for the Japanese language (reading, rough, joyful and sad) and for a single speaker, as the introduction of style-dependent questions in the tree-based state-tying process, which is analogous to a Maximum Likelihood-based clustering of the acoustic states;
- (Miyanaga et al., 2004), in a follow-up of the above work, have proposed a method to control the speaking style explicitly. The method is based on the adaptation of the HMM parameters through a multiple linear regression piloted by a style-control vector. The underlying idea is to use the speaking style as an auxiliary information for the adaptation of the model parameters, and the corresponding regression model can be determined by some EM re-estimation formulae. Some convincing demos of this method are available on the web, at the following address: `http://sp-www.ip.titech.ac.jp/demo/index.html`;
- (Tachibana et al., 2004) have successfully rendered a continuous range of speaking styles by applying some model interpolation techniques between the three speaker-dependent models of read, joyful and sad speech.

By supporting the quasi literal application of speaker adaptation techniques to emotional adaptation, the above-cited works suggest that the emotional speech data lends itself to the application of the same model subspace concepts as those underlying speaker adaptation techniques.

2.4.4 Relevance for PAVOQUE

From the above study, the following technical and strategic points can be considered in the perspective of PAVOQUE.

Models or units ? – GMMs provide an abstraction of the data, upon which an adaptation transform or a tree-based context clustering can be estimated. The resulting transform or clustering can be applied either to some actual units (voice transformation) or to the abstract model itself (GMM-based synthesis). Hence: either we could train GMMs, adapt and/or tree-cluster them and apply the trained transform to the real units, instead of the models, before concatenation; or, we could shift completely to the HMM paradigm as a joined transformation and speech generation paradigm.
On the one hand, HMM-based synthesis bypasses the unit transformation stage, but the parameter-generation process used to produce speech will require a significant effort to be fully understood and programmed. However, the HTS code is available, and it is built on top of HTK. This software seems fairly usable "out of the box" for HMM-based synthesis, provided an adequate database is available for training the models. We have installed the software under Linux without meeting any particular problem. This code has been reported to work within Festival and to have been used for the synthesis of English (Tokuda et al., 2002) and of Brazilian Portuguese (da S. Maia et al., 2003). Hence, it could be expected to interface gracefully with Mary and to be applicable to German. Apparently, HMM-based synthesis of German has already be attempted at the IKP in Bonn (Weiss et al., 2005), with the help of Tokuda's team from Nitech, and under the Bonn Open Synthesis System (BOSS), but with a very limited amount of training data which has led to questionable intelligibility. It is to be noted that although the version of HTS released in January 2007 includes the adaptation schemes advertised in the earliest of the related Japanese papers, the incorporation and release of the state-of-the-art in this open-source software package always requires a certain delay.
On the other hand, transformed natural units may bring a better perceptual quality, given that the weak point of HMM-based synthesis is the vocoder-like quality related to the artifical pulse or noise excitation that is used with the MLSA filters. This last problem seems about to be solved, given the results exposed in (Yoshimura et al., 2001) where a more elaborate excitation model is incorporated to the HMM paradigm. Nevertheless, this model still assumes the independent analysis of the excitation characteristics versus the speech spectrum, the latter being still represented by MFCCs and MLSA filtering. A possible innovation, as far as HMMs are concerned, could rely on using LPC-based features (possibly envelope-based) in place of MFCC features, to get a more consistent analysis/synthesis framework. In addition, a "second-stage" CELP-style short-term analysis could be used (Schroeder and Atal, 1985) to model the excitation spectrum in a more detailed way, and the adaptation of the excitation itself could be envisioned[2]. The validity of the excitation adaptation paradigm could be assessed with the help of the PAVOQUE database.

[2]As a matter of fact, the "1st stage" LPC can be dimensioned according to the expected number of formants (Atal and Hanauer, 1971) and in a way which would leave "enough" info in the residual spectrum for a possible adaptation of the residual itself.

To remain open, an ideal solution would be to design a synthesis framework where the Gaussian models and the units would be interchangeable, so that we could compare the performances of both approaches in a flexible way. An even better paradigm would consist in allowing HMM-based synthesis to compensate for missing natural units, in a sort of hybrid HMM/unit-based synthesis system (think of "variable quality synthesis", in the same way as "variable bit-rate coding").

Related database design issues – In any case: if we want to learn the emotional adaptation transform from the data, the dimensioning of the PAVOQUE database will play a very important role. Care should be taken in the dimensioning of three independent datasets:

- the training set, to train the units, the HMMs, the trees etc.;
- the development set, for an independent tuning of the training procedures;
- the testing set, to be able to include human sentences in the listening assessments and possibly to develop objective measures based on a contrastive measure between human and synthetic speech.

Care should be put into balancing the phonetic versus expressive variability in an informed way. For example, we should use the same sentences across emotional styles to facilitate the alignment which underlies the phoneme dependent mapping, but we should keep enough context variability to preserve the justification of the state-tying/context-clustering procedure. The fact that this database will be designed explicitely for the *training* of expressiveness models represents an important scientific asset, and should be advertised in related publications.

After a review of the works supporting the parameterisation and the transformation of voice quality as a spectral effect, let us now review the works dealing with the parameterisation and transformation of the prosody.

3 Parameterisation of Prosody

3.1 Modelling and modification of duration and F_0

3.1.1 Existing methods

The modelling of prosody is a whole area of research in itself. The most recent trend (van Santen et al., 2005) seems to build on Fujisaki's superpositional model (Fujisaki, 1988). In the framework of unit-selection based speech synthesis, the standard practice consists in relying on the target features as accurate-enough predictors of the prosody and durations, possibly with the help of ToBI-type features (Silverman et al., 1992). In the framework of the so-called "trainable" unit-based synthesis systems, such as the IBM system (Donovan and Eide, 1998; Pitrelli et al., 2006), the prosody and duration values are predicted explicitely and independently of the unit sequence, using a set of context-clustering CART-trees[3]. Alternately, in the framework of HMM-based synthesis, the prosody is predicted as an explicit value, but in correlation with the more detailed acoustic context, on the basis of the Gaussian Mixture Models underlying the HMMs (Yoshimura et al., 1999); the durations are themselves predicted from an external Gaussian

[3]The energy contour was also independently predicted, but the energy prediction has been found to play a negligible role on the synthesis quality. However, this was found when using an "equalised" database. We may want to revise this finding with our own emotional, non-equalised data.

model. Finally, when no model is available, human prosodic contours can be used as a template, with the help of DTW, a practice referred to as prosodic transplant (Verhelst and Borger, 1991; van Santen et al., 2005).
Following the modelling stage, the prosody of a series of concatenated speech units can be modified to correct some audible discontinuities, and/or to impose a completely different prosodic contour. For this kind of modification, the standard practice consists in shifting and/or duplicating some speech frames localised by some corresponding pitch marks. The much used Pitch-Synchronous OverLap Add (PSOLA) algorithm (Moulines and Charpentier, 1990) is an instance of such a paradigm, where the speech frames are plain waveforms; however, the OverLap-Add (OLA) paradigm can be generalised to any other form of coding of the speech frames, including spectral frames in the frequency domain (Moulines and Verhelst, 1995), or envelope-based LPC coefficients. For example, the MBROLA algorithm (Dutoit and Leich, 1993) is a particular instance of the OLA paradigm, where the shifts are independently operated across several spectral bands in order to minimise the phase distortions induced by the method. *In the literature studied for this report, we have not encountered any alternative to the OLA paradigm as far as prosody modifications are concerned.*

3.1.2 Notes about expression-relevant target modelling for F_0 and duration

The tree-based context clustering method seems applicable to emotion-based prosodic modelling: IBM trains separate prosody-prediction trees for separate expressions (Pitrelli et al., 2006).
In the Japanese HMM-based systems, emotion-based MLLR adaptation manages the pitch and duration, jointly with the spectral shapes, in the framework of speaker-to-speaker voice transformation (Tamura et al., 2001). In a follow-up, speaking style oriented work, HMM state tying on the basis of a tree-based context clustering including speaking style labels has been performed separately for the spectral, F_0 and duration parts of the model (Yamagishi et al., 2003). The results show good emotion classification performances and a fair synthesis quality (demos available on: `http://sp-www.ip.titech.ac.jp/research/demo/`). On the basis of this tree-based speaking style modelling method, (Miyanaga et al., 2004) have defined a speaking style control technique based on a regression matrix estimated from the tree-clustered states, with interesting preliminary results in terms of definition of separate control parameters acting on style features that emerge automatically from data analysis. *This works illustrate the close integration of state-tying and MLLR transformation, which is a major asset of HMMs as a supporting model for emotional transformations.*

3.2 Technical issues and strategy

3.2.1 Pitch analysis

The methods available for the automatic extraction of pitch marks or pitch frequency are numerous and diverse. For example, see the review in (Bagshaw, 1994) and the numerous references that can be found in the ICASSP or Interspeech conferences. Since our primary goal is not to contribute new methods to this field, our technological choices should rather be based on the practical comparison of some standard open-source implementations. These include:

- the `pitchmark` module of the Edinburgh Speech Tools, based on a rather crude time-domain filtering method initially designed for the analysis of laryngograph signals;

- in ESPS: the `get_f0` module (re-used in Wavesurfer), based on the normalised cross correlation function and dynamic programming, and described in (Talkin et al., 1995); or the `epochs` module, initially targeted at epoch detection from LPC residuals, and described in (Talkin, 1989; Talkin and Rowley, 1990). Relevant man pages are available with ESPS;

- the pitch extraction algorithm embedded in Praat (`http://www.praat.org/`), which results from an autocorrelation-based method and is described in (Boersma, 1993).

These three environments have been compiled and installed at DFKI without difficulty, but the systematic comparison of their performances, together with an assessment of the implications of their interfacing with Mary, has yet to be performed.

An additional alternative, in relation to the sinusoidal analysis issues exposed in section 2.3.2, would be to study the potentialities of Matching Pursuit (Krstulović and Gribonval, 2006) for pitch extraction (or "pitch matching"). It would indeed be interesting to study the behaviour of the algorithm when analysing speech with dictionaries of harmonic atoms, using adequate constraints in terms of window lengths, window shifts, human-compatible F0 range and constraints on the authorised atom overlap, and with the possible addition of a contour-extraction post-processing. This would represent some possible innovation, having the additional advantage of a better integration with high-quality envelope-based LPC estimation or Harmonics plus noise models, as suggested in section 2.3.2.

3.2.2 Duration analysis

The common practice for duration analysis seems to rely either on the manual labelling of the speech database, or on the determination of the state durations from an automatic HMM alignment (using Sphinx or HTK). We have not encountered any alternative paradigm to this in the speech synthesis literature studied for the needs of this report.

3.2.3 Tree-based context clustering

In the above-cited works, the prosody-prediction methods make use of tree-based context clustering. As far as software is concerned, we currently only dispose of EST/wagon to train the CART trees. This software is coded in C++ and imposes severe restrictions on the tree topology, the purity measures available to train them and the I/O data formats: its lack of flexibility will become an obstacle in the long term. Apart from the tree clustering algorithms included in HTK for state-clustering, following Julian Odell's PhD (Odell, 1995), it is not clear which software has been used by other researchers (e.g., IBM/Donovan) to train their trees for prosody or duration prediction. Public domain tree-training software is available in the form of the C4.5 legacy code from Ross Quinlan (`http://www.rulequest.com/Personal/`), or Howard Hamilton's C code (`http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html`), or the Java WEKA package (`http://www.cs.waikato.ac.nz/ml/weka/`). A practical assessment of these alternatives, in comparison to EST/Wagon, is yet to be performed.
Concerning prosodic models, the span of our historical background in the domain incurs the risk of getting lost if we try to go too deep into the modelling direction. A reasonable first step would therefore consist in focusing on the development of a flexible pitch/duration modification paradigm at the software implementation level, based on a generic OLA paradigm, so that we could provide a flexible support for the testing of some external prosodic prediction models that could be introduced at a later stage.

4 Conclusion and road map

The present report has started by showing that voice quality transformations based on acoustical models were appearing more tractable than transformations based on production models, in particular those relying on some explicit glottal waveform models. As a matter of fact, the evolution of speaker modification techniques has seen Gaussian-based acoustic modelling emerge as the primary support for the state of the art, because a flexible statistical model is needed to abstract the acoustical realizations of the speech units by a limited number of parameters that are still able to account for some variability, and because the speech transformations may need to be defined in a class-dependent way (e.g., different transformations for different phonemes), across classes which should preferably automatically emerge from the data (through unsupervised learning).
Besides, a range of Gaussian-based speaker adaptation techniques has been extensively developed for speech and speaker recognition, and their application to HMM-based voice transformations has given promising results. Then, sets of HMMs or Gaussian models adapted to distinct speech classes can also be used to define a model space, where more recent methods such as model interpolation or eigenvoices can be applied to obtain a more explicit control of speaker-specific or style-specific variations.
Finally, HMM-based synthesis techniques readily incorporate some tree-based prosody and duration modelling techniques inherited from unit-based or trainable synthesis, in a framework which unifies the modelling of prosody and voice quality. This framework also provides a paradigm for interpolating unseen speech units.
For all these reasons, supported by concrete results from other authors, and in the perspective of the goals of the PAVOQUE project, we recommend the use of HMM-based synthesis as a core technological framework to support research about the parameterisation of voice quality and prosody in view of expressive speech synthesis.

References

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. In *Proc. ICASSP'88*, pages 655–658, New York.

Airas, M., Pulakka, H., Bäckström, T., and Alku, P. (2005). A toolkit for voice inverse filtering and parametrisation. In *Proc. Interspeech'05*, pages 2145–2148, Lisbon, Portugal.

Akande, O. O. and Murphy, P. J. (2005). Estimation of the vocal tract tranfer function with application to glottal wave analysis. *Speech Communication*, 46:15–36.

Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118.

Alku, P. (2003). Parameterisation methods of the glottal flow estimated by inverse filtering. In *Proc. VOQUAL'03*, pages 81–87, Geneva.

Arslan, L. M. (1999). Speaker tranformation algorithm using segmental codebooks (STASC). *Speech Communication*, 28:211–226.

Atal, B. and Hanauer, S. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, 50(2):637–655.

Bagshaw, P. (1994). *Automatic prosodic analysis for computer aided pronunciation teching.* PhD thesis, University of Edinburgh.

Baudoin, G. and Stylianou, Y. (1996). On the transformation of the speech spectrum for voice conversion. In *Proc. ICSLP'96*, Philadelphia, PA, USA.

Black, A. W. and Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. In *Proc. Eurospeech'97*, pages 601–604, Rhodes, Greece.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and harmonics-to-noise ratio of a sampled sound. *IFA Proceedings*, 17:97–110.

Cappé, O., Laroche, J., and Moulines, E. (1995). Regularized estimation of cepstrum envelope from discrete frequency points. In *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk.

Childers, D. G. (1995). Glottal source modeling for voice conversion. *Speech Communication*, 16:127–138.

da S. Maia, R., Zen, H., Tokuda, K., Kitamura, T., and ResendeJr, F. (2003). Towards the development of a brazilian portuguese text-to-speech system based on HMM. In *Proc. Eurospeech'03*, pages 2465–2468, Geneva.

d'Alessandro, C. and Doval, B. (1998). Experiments in voice quality modification of natural speech signals: the spectral approach. In *Proc. 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, Jenolan Caves House, Blue Mountains, NSW, Australia.

d'Alessandro, C. and Doval, B. (2003). Voice quality modification for emotional speech synthesis. In *Proc. Eurospeech'03*, pages 1653–1656.

Donovan, R. and Eide, E. (1998). The IBM trainable speech synthesis system. In *Proc. ICSLP'98*, Sydney, Australia.

Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text to speech synthesis based on a MBE re-synthesis of the segments database. *Speech Communication*, 13.

Fant, G. (1970). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G., Liljentcrats, J., and Lin, Q. (1985). A four parameter model of glottal flow. In *Quarterly Progress and Status Report*, number 4 in STL-QPSR, pages 1–13. KTH, Stockholm, Sweden.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura, O., editor, *Vocal Physiology: Voice Production*, pages 347–355. Raven, New York.

Galas, T. and Rodet, X. (1991). Generalized functional approximation for source-filter system modeling. In *Proc. Eurspeech'91*, pages 1085–1088, Genova, Italy.

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. ICASSP'83*, pages 93–96, Boston.

Jaroudi, A. E. and Makhoul, J. (1991). Discrete all-pole modeling. *IEEE Trans. on Signal Processing*, 39(2):411–423.

Kain, A. and Macon, M. W. (1998). Spectral voice conversion for text-to-speech synthesis. In *Proc. ICASSP'98*, volume 1, pages 285–288.

Kain, A. and Macon, M. W. (2001). Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction. In *Proc. ICASSP'01*.

Krstulović, S. and Gribonval, R. (2006). Mptk: Matching pursuit made tractable. In *Proc. ICASSP'06*, Toulouse, France.

Kuwabara, H. and Sagisaka, Y. (1995). Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communication*, 16:165–173.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.

Liu, J., Beaudoin, G., and Chollet, G. (1992). Studies of glottal excitation and vocal tract parameters using inverse filtering and a parameterized input model. In *Proc. ICSLP'92*, pages 1051–1054, Banff, Alberta, Canada.

Lugger, M., Yang, B., and Wokurek, W. (2006). Robust estimation of voice quality parameters under real world disturbances. In *Proc. ICASSP'06*, pages 1097–1100.

Macchi, M., Altom, M. J., Kahn, D., Singhal, S., and Spiegel, M. (1993). Intelligibility as a function of speech-coding method for template-based speech synthesis. In *Proc. Eurospeech'93*, pages 893–896, Berlin, Germany.

Macon, M. (1996). *Speech Synthesis based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology.

Macon, M. W. and Clements, M. A. (1999). An enhanced ABS/OLA sinusoidal model for waveform synthesis in TTS. In *Proc. Eurospeech'99*, volume 5, pages 2327–2330.

Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (1997). Voice characteristics conversion for HMM-based speech synthesis. In *Proc. ICASSP'97*, pages 1611–1614.

Milenkovic, P. (1993). Voice source model for continuous control of pitch period. *Journal of the Acoustical Society of America*, 93(2):1087–1096.

Miyanaga, K., Masuko, T., and Kobayashi, T. (2004). A style control technique for HMM-based speech synthesis. In *Proc. ICSLP'04*, Jeju, Korea.

Mokbel, C., Pachès-Leal, P., Jouvet, D., and Monné, J. (1994). Compensation of telephone line effects for robust speech recognition. In *Proc. ICSLP'94*, pages 987–990, Yokohama, Japan.

Mokhtari, P., Pfitzinger, H. R., and Ishi, C. T. (2003). Principal components of glottal waveforms: towards parameterisation and manipulation of laryngeal voice quality. In *Proc. VOQUAL'03*, Geneva.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5):453–467.

Moulines, E. and Verhelst, W. (1995). Time-domain and frequency-domain techniques for prosodic modification of speech. In Kleijn, W. and Paliwal, K., editors, *Speech Coding and Synthesis*, chapter 15, pages 519–555. Elsevier Science B.V.

Odell, J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, Queen's College, University of Cambridge.

Oppenheim, A. W. and Schafer, R. W. (1968). Homomorphic analysis of speech. *IEEE Transactions on Audio and Electracoustics*, AU-16(2):221–226.

Paliwal, K. (1995). Interpolation properties of linear prediction parametric representations. In *Proc. Eurospeech'95*, pages 1029–1032, Madrid, Spain.

Pitrelli, J., Bakis, R., Eide, E., Fernandez, R., Hamza, W., and Picheny, M. (2006). The IBM expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1099–1108.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Trans. on Speech and Audio Processing*, 2(4):639–643.

Roehling, S., MacDonald, B., and Watson, C. (2006). Towards expressive speech synthesis in english on a robotic platform. In *Proc. 11th Australasian International Conference on Speech Science and Technology*, Auckland, New Zealand. Univ. of Auckland.

Schafer, R. W. and Rabiner, L. R. (1990). Digital representations of speech signals. In Waibel, A. and Lee, K. F., editors, *Readings in Speech Recognition*, pages 49–64. Morgan Kaufmann.

Schröder, M. (2001). Emotional speech synthesis: A review. In *Proc. Eurospeech'01*, Scandinavia.

Schröder, M. (2004). *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. PhD thesis, Institut für Phonetik , Universität des Saarlandes. Phonus no.7.

Schroeder, M. R. and Atal, B. S. (1985). Code-excited linear prediction (celp): High-quality speech at very low bit rates. In *Proc. ICASSP'85*, pages 937–940.

Schultheiss, M. and Lacroix, A. (1989). On the performance of CELP algorithms for low rate speech coding. In *Proc. ICASSP'89*, volume 1, pages 152–155.

Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2002). Eigenvoices for HMM-based speech synthesis. In *Proc. ICSLP'02*, Denver, Colorado.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., and Hirschberg, J. (1992). Tobi: A standard scheme for labeling prosody. In *Proc. ICSLP'92*, Banff.

Stylianou, Y., Cappé, O., and Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Trans. on Speech and Audio Processing*, 6(2):131–142.

Stylianou, Y., Laroche, J., and Moulines, E. (1995). High-quality speech modification based on a harmonic + noise model. In *Proc. Eurospeech'95*, pages 451–454, Madrid, Spain.

Tachibana, M., Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2004). HMM-based speech synthesis with various speaking styles using model interpolation. In *Proc. Speech Prosody 2004*, Nara, Japan.

Talkin, D. (1987). Speech formant trajectory estimation using dynamic programming with modulated transition costs. Technical report, Bell Labs.

Talkin, D. (1989). Voicing epoch determination with dynamic programming. *J. Acoust. Soc. Amer.*, 85(Supplement 1).

Talkin, D., Kleijn, W., and Paliwal, K. (1995). A robust algorithm for pitch tracking (rapt). In *Speech Coding and Synthesis*, pages 495–518. Elsevier.

Talkin, D. and Rowley, J. (1990). Pitch-synchronous analysis and synthesis for tts systems. In Benoit, C., editor, *Proceedings of the ESCA Workshop on Speech Synthesis*, Gieres, France.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (1998). Speaker adaptation for HMM-based speech synthesis using MLLR. In *Proc. 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, Blue Mountains, Australia.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2001). Text-to-speech synthesis with arbitrary speaker's voice from average voice. In *Proc. Eurospeech'01*, Scandinavia.

Tokuda, K., Zen, H., and Black, A. W. (2002). An HMM-based speech synthesis system applied to english. In *Proc. IEEE Speech Synthesis Workshop*, Santa Monica, California.

Turk, O. and Arslan, L. M. (2006). Robust processing techniques for voice conversion. *Computer Speech and Language*, 20:441–467.

Turk, O., Schröder, M., Bozkurt, B., and Arslan, L. (2005). Voice quality interpolation for emotional text-to-speech synthesis. In *Proc. Interspeech'05*, pages 797–800, Lisbon, Portugal.

Valbret, H., Moulines, E., and Tubach, J. (1992). Voice transformation using PSOLA technique. *Speech Communication*, 11:175–187.

van Santen, J., Kain, A., Klabbers, E., and Mishra, T. (2005). Synthesis of prosody using multi-level unit sequences. *Speech Communication*, 46:365–375.

Verhelst, W. and Borger, M. (1991). Intra-speaker transplantation of speech characteristics, an application of vocoding techniques and dtw. In *Proc. Eurospeech'91*, pages 1319–1322, Genova, Italy.

Villavicencio, F., Röbel, A., and Rodet, X. (2006). Improving LPC spectral envelope extraction of voiced speech by true envelope estimation. In *Proc. ICASSP'06*, volume I, pages 869–872, Toulouse, France.

Weiss, C., Maia, R. D. S., Tokuda, K., and Hess, W. (2005). Low resource hmm-based speech synthesis applied to german. In *Proc. 16th Conference on Electronic Speech Signal Processing, joint with the 15th Czech-German Speech Processing Workshop (ESSP2005)*, Prag, Czech Republic.

Wokurek, W. and Pützer, M. (2003). Automated corpus based spectral measurement of voice quality parameters. In *Proc. 15th ICPhS*, pages 2173–2176, Barcelona.

Woodland, P. (2001). Speaker adaptation for continuous density hmms: a review. In *Proc. ITRW on Adaptation Methods for Speech Recognition*, pages 11–19, Sophia Antipolis.

Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2003). Modeling of various speaking styles and emotions for HMM-based speech synthesis. In *Proc. Eurospeech'03*, pages 2461–2464, Geneva, Switzerland.

Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., and Kitamura, T. (1997). Speaker interpolation in HMM-based speech synthesis system. In *Proc. Eurospeech'97*, Rhodos, Greece.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech'99*, volume 5, pages 2347–2350, Budapest, Hungary.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2001). Mixed excitation for hmm-based speech synthesis. In *Proc. Eurospeech'01*, Scandinavia.

Young, S., Odell, J., and Woodland, P. (1994). Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Workshop on Human Language Technology*, pages 307–312.

A review of state-of-the-art speech modelling methods for the parameterisation of expressive synthetic speech

07-02
Technical Memo

Sacha Krstulović