LISTA D3.1 – Speaker and spectral adaptation
LISTA D3.2 – Prosodic adaptation

Junichi Yamagishi, Cassia Valentini-Botinhao, Cassie Mayo,
Simon King, Julian Villegas, Gustav Eje Henter, Yannis Stylianou

April 20, 2012

Contents

1 Introduction

This deliverable starts with an introduction to Hidden Markov model-based speech synthesis and in particular the variety of existing adaptation methods available. This part of the deliverable is provided as background material and is not original work conducted in LISTA. However, it will be useful in understanding some of the work reported in the remainder of the deliverable, which employs these adaptation techniques.

Then, the work within LISTA is reported, broken down by task. Where the work has been published, descriptions are kept brief and the corresponding publications are cited.

1.1 Hidden Markov model (HMM)-based speech synthesis

Current text-to-speech synthesis systems can generate high quality speech. However it is still not straightforward to synthesize speech with various voice characteristics such as speaking style, or the Lombard effect. To achieve these voice characteristics with conventional concatenative unit selection systems, a large amount of speech data for each individual voice characteristic is necessary; hence, unit selection is not a practical, scalable or cost efficient approach.

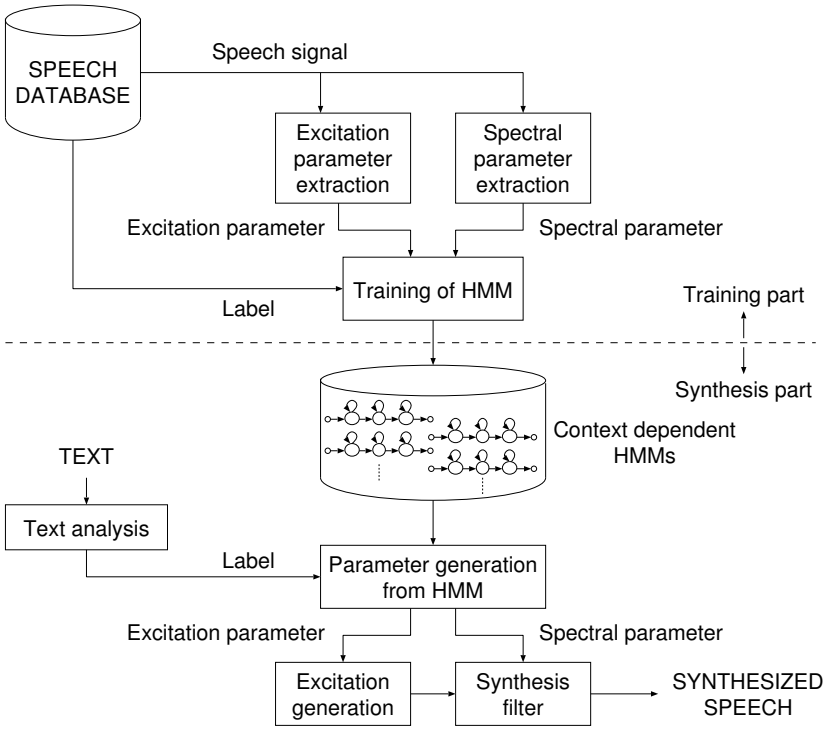

Figure 1: HMM-based speech synthesis system.

In order to create more flexible, adaptable and controllable text-to-speech synthesis systems, HMM-based speech synthesis was proposed (Yoshimura et al., 1999). Figure 1 shows an overview of a typical system of this kind. In the training part, spectrum and excitation parameters are extracted from a speech database and modeled by context-dependent phoneme HMMs. In the synthesis part, these context-dependent HMMs are concatenated according to the text to be synthesized. Then, spectrum and excitation parameters are generated from the HMM by using a speech parameter generation algorithm

(Tokuda et al., 2000). Finally, the excitation generation module and synthesis filter module synthesize the speech waveform using the generated excitation and spectrum parameters. The speech feature extraction stage and the later waveform generation stage are essentially the two halves of a vocoder, and the statistical models inbetween operate in the vocoder parameter domain. This is worth highlighting in the context of LISTA, since it suggests multiple domains in which speech modification could be achieved, including the waveform, vocoder parameters, or the models themselves.

The training part performs maximum likelihood estimation of HMM parameters by using the Baum-Welch algorithm. This process is very similar to the one used for speech recognition, the main difference being that both spectrum (e.g., mel-cepstral coefficients (Fukada et al., 1992) and their dynamic features) and excitation (e.g., $\log F_0$ and its dynamic features) parameters are modeled by a set of multi-stream context-dependent HMMs. Another difference is that both segmental (i.e., phonetic) and supra-segmental (e.g., prosodic) contexts are taken into account. To model fixed-dimensional speech features, such as mel-cepstral coefficients, single multi-variate Gaussian distributions are typically used. However, it is difficult to apply discrete or continuous distributions to model variable-dimensional speech parameters, such as $\log F_0$ sequences with voiced (dimension = 1) and unvoiced (dimension = 0) regions. For modeling $\log F_0$ sequences, HMM-based speech synthesis system most commonly adopts "multi-space probability distributions" (Tokuda et al., 2002) as stream-output probability distributions. Each HMM also has its state-duration probability distribution called the 'semi-Markov probability' to model the temporal structure of speech (Zen et al., 2004). This is typically a Gaussian distribution. Each of the spectrum, excitation, and duration model parameters are clustered separately by using decision trees (Odell, 1995), because the way in which each of them depends on context will vary. Overall, the system models the spectrum, excitation, and duration in a unified framework.

The synthesis part of the system is shown in the lower part of Fig. 1. It first converts a given text to be synthesized to a sequence of context-dependent labels. According to the label sequence, a sentence-level HMM is constructed by concatenating context-dependent HMMs. The duration of each state is determined so as to maximize its probability based on its state-duration probability distribution. Then a sequence of speech parameters including spectral and excitation parameters is determined so as to maximize their probability according to the speech parameter generation algorithm (Tokuda et al., 2000). The main feature of the algorithm is the use of dynamic features; by imposing the relationship between static and dynamic features, the most probable speech parameter trajectory is constrained to be a realistic, smoothly varying sequence. Finally, a speech waveform is resynthesized directly from the generated spectral and excitation parameters by using a speech synthesis filter, such as the mel-log spectral approximation filter for mel-cepstral coefficients or an all-pole filter for linear prediction-based spectral parameters coefficients.

1.2 Speaker adaptation in HMM-based speech synthesis

The main advantage of HMM-based speech synthesis is its flexibility in changing voice characteristics and speaking styles – we can easily do this by transforming the model parameters. There are currently four principal techniques to accomplish this: adaptation, interpolation, eigenvoices, and regression techniques. The adaptation technique is evaluated in Section 2 and the regression technique is evaluated in Section 3.

Speaker adaptation may be employed to transform existing speaker-independent acoustic models to match a target speaker using a very small amount of speech data (Yamagishi et al., 2009). This method starts with an "average voice model" and uses model adaptation techniques drawn from speech recogni-

tion such as maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Woodland, 2001), to adapt the speaker independent HMMs to a new speaker or to a new speaking style.

This adaptation allows text-to-speech synthesizers to build a target voice using much smaller amounts of training data than previously required. Prior to this, the development of a new voice required many hours of carefully annotated speech recordings from a single speaker. Speaker adaptive HMM-based synthesis requires as little as 5–7 minutes of recorded speech from a target speaker in order to generate a personalized synthetic voice (Yamagishi et al., 2009). The main adaptation techniques used in HMM-based speech synthesis are similar to those used in ASR and include maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994), MLLR (Leggetter and Woodland, 1995) and vocal tract length normalisation (VTLN) (McDonough, 2000; Kim et al., 2004).

In the following subsections, we overview these adaptation techniques and explain how they are related in the spectral domain, which is the focus of D3.1.

1.3 VTLN

The main components of VTLN are a warping function, a warping factor and an optimization criterion. Typically, the warping function has only a single variable α as the warping factor, which is representative of the ratio of the vocal tract length of a speaker to an average vocal tract length. In ASR, where a mel or bark spaced filter bank is used, the warping function tends to be linear or piecewise-linear, and is normally applied directly to the filter-bank.

By contrast, feature extraction for TTS tends not to use a filter-bank analysis as it renders signal reconstruction difficult. Rather, the feature commonly used in TTS is the mel-generalized cepstrum (Tokuda et al., 1994), which makes use of a bilinear transform to achieve the frequency warp[1]. Since the mel-generalized cepstrum already includes a bilinear transform, then bilinear transform-based VTLN proposed by Pitz and Ney (Pitz and Ney, 2005) can be implemented as a zero-overhead modification of the spectral representation.

The bilinear transform of a simple first-order all-pass filter with unit gain leads to a warping of the frequency ω into $\tilde{\omega}$ in the complex z-domain as follows:

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{1}$$

where $z^{-1} = e^{-j\omega}$, $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$, and α is the warping factor. We define the m-th mel-cepstral coefficient, that is, frequency warped cepstrum, $\tilde{c}_m$ as

$$\tilde{c}_m = \frac{1}{2\pi j} \oint_C \log X(\tilde{z})\, \tilde{z}^{m-1} d\tilde{z} \tag{2}$$

$$\log X(\tilde{z}) = \sum_{m=-\infty}^{\infty} \tilde{c}_m\, \tilde{z}^{-m} \tag{3}$$

Since the frequency warping is $X(\tilde{z}) = X(z)$, we have a linear transformation in the cepstral domain

[1]Spectral analysis in the mel-generalized cepstrum also uses a generalized logarithmic function, which has the effect of varying the analysis between an all-pole and a cepstral model, according to a second parameter.

c_k:

$$\tilde{c}_m = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi j} \oint_C \tilde{z}^{-k} z^{m-1} d\tilde{z}\, c_k \tag{4}$$

$$= \sum_k A_{mk}(\alpha)\, c_k \tag{5}$$

where $A_{mk}(\alpha)$ is the m-th row k-th column element of the warping matrix $\boldsymbol{A}_\alpha$ consisting of the warping factor α and the Cauchy integral formula yields (Pitz and Ney, 2005):

$$A_{mk}(\alpha) = \frac{1}{2\pi j} \oint_C \tilde{z}^{-k} z^{m-1} d\tilde{z} \tag{6}$$

$$= \frac{1}{2\pi j} \oint_C \left(\frac{z-\alpha}{1-\alpha z} \right)^{-k} z^{m-1}\, d\tilde{z} \tag{7}$$

$$= \frac{1}{(k-1)!} \sum_{n=\max(0,k-m)}^{k} \binom{k}{n} \times \frac{(m+n-1)!}{(m+n-k)!} (-1)^n \alpha^{2n+m-k}. \tag{8}$$

If we truncate the original and warped mel-cepstral coefficients at K-th and M-th dimensions, we may represent VTLN in the linear transformation form below,

$$\boldsymbol{x}_\alpha = \boldsymbol{A}_\alpha \boldsymbol{x} \tag{9}$$

where $\boldsymbol{x}_\alpha = (\tilde{c}_1, \cdots, \tilde{c}_M)^\top$ and $\boldsymbol{x} = (c_1, \cdots, c_K)^\top$. The transform may also be directly applied to the dynamic features of the cepstra. The transformation matrix is block diagonal with repeating $\boldsymbol{A}_\alpha$ matrix.

There are several ways to estimate the warping factor α. For example, the maximum likelihood criterion may be adopted α (Lee and Rose, 1998).

$$\widehat{\alpha}_s = \operatorname*{argmax}_\alpha P(\boldsymbol{x}_{\alpha_s} \mid \lambda,\, \alpha_s) \tag{10}$$

where $\boldsymbol{x}_{\alpha_s}$ represents features warped with the warping factor α_s for speaker s; λ represents average voice models, and $\widehat{\alpha}_s$ represents the optimal warping factor for speaker s.

1.4 CMLLR and CSMAPLR

If we remove all the explicit constraints specified by the warping factor α from the linear matrix $\boldsymbol{A}_\alpha$ and add a bias term $\boldsymbol{b}$ to (9), this becomes a simple affine transform of the spectral feature $\boldsymbol{x}$.

If we estimate the affine transform of observation vectors including the spectral feature $\boldsymbol{x}$ as a part of HMM parameters, this is called constrained maximum likelihood linear regression (CMLLR) (Gales,

1998) and the likelihood function of HMM parameters λ including the affine transform can be written as

$$P(\boldsymbol{X}|\lambda) = \sum_{\boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{x}_t), \tag{11}$$

$$\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_T^\top]^\top, \tag{12}$$

$$b_i(\boldsymbol{x}_t) = |\boldsymbol{A}_i| \mathcal{N}(\widehat{\boldsymbol{x}}_t;\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{13}$$

$$= |\boldsymbol{A}_i| \mathcal{N}(\boldsymbol{A}_i \boldsymbol{x}_t + \boldsymbol{b}_i;\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{14}$$

$$= |\boldsymbol{A}_i| \mathcal{N}(\boldsymbol{W}_i \boldsymbol{\xi}_t;\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{15}$$

$$\boldsymbol{W}_i = [\boldsymbol{A}_i,\ \boldsymbol{b}_i] \tag{16}$$

$$\boldsymbol{\xi}_t = [\boldsymbol{x}_t^\top, 1]^\top \tag{17}$$

$$\widehat{\boldsymbol{x}}_t = \boldsymbol{A}_i \boldsymbol{x}_t + \boldsymbol{b}_i \tag{18}$$

$$= \boldsymbol{W}_i \boldsymbol{\xi}_t \tag{19}$$

where where π_j and a_{ij} represent initial state probability and state transition probability of an HMM; $b_i(\cdot)$ is the state observation probability density function (PDF) for a state i of the HMM; $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ is the state sequence for the observation sequence $\boldsymbol{X}$; $\boldsymbol{A}_i \in \mathcal{R}^{K+1_X \times (K+1)}$ is the affine transform matrix for state i.

CSMAPLR (Yamagishi et al., 2009) is a robust framework to estimate the CMLLR transforms based on the SMAP criterion (Shiohan et al., 2002; Shinoda and Lee, 2001):

$$\widehat{\boldsymbol{W}}_i = [\widehat{\boldsymbol{A}}_i,\ \widehat{\boldsymbol{b}}_i] = \underset{\boldsymbol{W}_i}{\operatorname{argmax}}\, P(\boldsymbol{X} \mid \lambda)\, P(\boldsymbol{W}_i) \tag{20}$$

where $\boldsymbol{W}_i$ refers to the set of CMLLR transforms for the state i. $P(\boldsymbol{X} \mid \lambda)$ is a likelihood function for $\boldsymbol{W}_i$ and $P(\boldsymbol{W}_i)$ is a prior distribution of the transform $\boldsymbol{W}_i$. Matrix variate normal distributions are used as the prior distribution $P(\boldsymbol{W})$:

$$P(\boldsymbol{W}) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{W} - \boldsymbol{H})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{W} - \boldsymbol{H}) \boldsymbol{\Psi}^{-1}\right] \tag{21}$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{L \times L}$, $\boldsymbol{\Psi} \in \mathbb{R}^{(L+1) \times (L+1)}$ and $\boldsymbol{H} \in \mathbb{R}^{L \times (L+1)}$ are the hyperparameters of the prior distribution.

In the SMAP criterion, the tree structures of the distributions effectively control these hyperparameters. The whole adaptation data is used to estimate a global transform at the root node of the tree based on the ML criterion and it is propagated to the child nodes as a hyperparameter $\boldsymbol{H}$. The transforms at each child node are estimated using the corresponding adaptation data and hyperparameters propagated with the MAP criterion. This process is continued recursively from the root node to all the leaf nodes of the tree structure.

In the CSMAPLR estimation, the hyperparameter $\boldsymbol{\Psi}$ is fixed to the identity matrix and $\boldsymbol{\Omega}$ to a scaled identity matrix, $\boldsymbol{\Omega} = \tau_b \boldsymbol{I}_L$. τ_b is a positive scalar that controls the scale factor for the prior propagation and $\boldsymbol{I}_L$ is $L \times L$. The hyperparameter of the prior distribution $\boldsymbol{H}$ is normally set to $\widehat{\boldsymbol{W}}_i$ of the parent node apart from the root node of the tree structure, which use an identity matrix, that is, no occupancy and statistics smoothing.

1.5 MAP adaptation on the top of linear transforms

It is possible to combine linear regression and MAP adaptation (Digalakis and Neumeyer, 1996; Chien et al., 1997). In the previous speaker adaptation method using linear regression, there is an implicit assumption that the target model can be approximated by a piecewise linear regression of the average voice model. By additionally applying MAP adaptation onto the model transformed by the linear regression, it is possible to further refine the distributions that have sufficient data (Fig. 2).

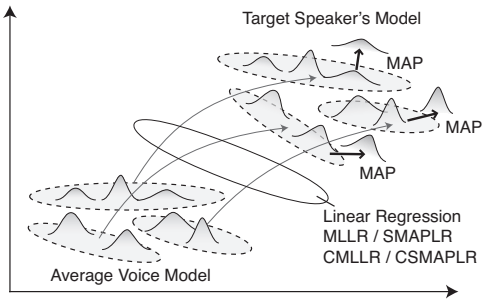

Figure 2: A combined algorithm of linear regression and MAP adaptation.

The MAP adaptation of mean vectors of the Gaussian pdfs adapted by the CSMAPLR transforms $[\widehat{\boldsymbol{A}}_i, \widehat{\boldsymbol{b}}_i]$ can be estimated as follows:

$$\widehat{\boldsymbol{\mu}}_i = \frac{\upsilon_b \, \boldsymbol{\mu}_i + \sum_{t=1}^{T} \gamma_t(i) \, \widehat{\boldsymbol{x}}_t}{\upsilon_b + \sum_{t=1}^{T} \gamma_t(i)} \tag{22}$$

$$\widehat{\boldsymbol{x}}_t = \widehat{\boldsymbol{A}}_i \boldsymbol{x}_t + \widehat{\boldsymbol{b}}_i \tag{23}$$

where $\boldsymbol{\mu}_i$ is a mean vector of the state output distribution of the average voice model, and $\widehat{\boldsymbol{x}}_s$ is a linearly transformed observation vector using CSMAPLR adaptation. $\gamma_t(i)$ is a state-occupancy probability of state i at time t. υ_b is a positive hyperparameter of the prior distribution for the state output distributions, respectively. Similarly we can combine any other linear regression algorithm with MAP adaptation. As the amount of adaptation data increases and the number of distributions having sufficient data increases, and so the adaptation performance improves from coarse to fine.

1.6 Extrapolation of the adapted model

It is further possible to enhance voice characteristics by applying an HMM extrapolation technique (Yoshimura et al., 2000) onto the MAP-adapted model. The final extrapolated adapted mean vector $\widetilde{\boldsymbol{\mu}}$ and covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ at state i are calculated as follows:

$$\widetilde{\boldsymbol{\mu}}_i = w \, \widehat{\boldsymbol{A}}_i^{-1} (\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{b}}_i) + (1 - w) \boldsymbol{\mu}_i \tag{24}$$

$$\widetilde{\boldsymbol{\Sigma}}_i = w^2 \, \widehat{\boldsymbol{A}}_i^{-1} \widehat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{A}}_i^{-1^\top} + (1 - w)^2 \boldsymbol{\Sigma}_i \tag{25}$$

where $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$ are the MAP-adapted mean vector and covariance matrix and $[\widehat{\boldsymbol{A}}_i, \widehat{\boldsymbol{b}}_i]$ are the CSMAPLR transforms. $\cdot^{-1}$ and $\cdot^\top$ represent matrix inverse and transpose, respectively. $\boldsymbol{\mu}_i$ is, again, the mean vector of a state output distribution of the average voice model. w is an interpolation ratio between adapted and average voices. All the Gaussian pdfs for all the acoustic features and durations are extrapolated in a similar way.

1.7 F0 and duration adaptation

One of the advantages of the adaptation techniques mentioned here is that these techniques (apart from VTLN) can be applied to any type of acoustic features, including F0 and duration (Yamagishi and Kobayashi, 2007). With regard to the LISTA goals, this means that speaker adaptation is able to mimic the key aspects of Lombard speech, which typically has longer vowel duration and higher F0 as well as flatter spectra.

2 Task T3.1 Direct use of speaker adaptation for the intelligibility improvement

Given the background we have just presented on speaker adaptation, and recalling that adaptation doesn't have to be to another speaker, it can just as well be to another speaking style of the same speaker, we now look at the first approach we took to synthesising speech intended to be heard in the presence of additive noise. This method is very simple, yet expected to work well. We used recordings of human talkers made in noisy conditions, who are performing natural and appropriate listener adaptation, because of the talking situation they find themselves in. Given such recordings, the speaker adaptation methods above can easily be applied to the speech synthesis models. However, this approach is limited by the requirement for recordings of human talkers, which may need to be from closely-matched environments: the synthesiser will "imitate" the human talkers' speech, but there will be no explicit control and therefore this method will probably not generalise to different listening situations for which no speech recordings exist. It certainly does not generalise directly to other speakers for whom we only have normal speech recordings.

Note that, in addition to the results presented below, a larger listening test was conducted to compare this technique with other methods developed in the LISTA project. This test was called the "Hurricane Challenge" and is described in deliverable D5.1.

2.1 Experiment on the Nick corpus

To build the voices used in this evaluation, we used two different datasets recorded by the same British male speaker "Nick": normal (plain, read-text) speech data and Lombard speech. The Lombard dataset was recorded while the speaker listened to speech-modulated noise based on another male speaker (Dreschler et al., 2001) played over headphones at an absolute value of 84 dBA.

Normal voice N was created from a high quality average voice model adapted to 2803 sentences of the normal speech database, corresponding to three hours of material. The speaker adaptation algorithm used was CSMAPLR+MAP. We decided to use an average voice model rather than building a speaker-dependent voice because the normal speech dataset was not phonetically balanced.

Lombard voice L was further adapted using 780 sentences from the Lombard speech dataset, corresponding to 53 minutes of recorded material. Again, the reason for using adaptation was the lack of phonetic balance in the speech dataset. Voice L-E is a version of voice L where we extrapolated the adaptation with a factor of 1.2.

Table 1 provides an acoustic analysis of the voices – average duration of speech and pauses, average spectral tilt, and F0 – across all sentences used in the listening test for the normal (N) and Lombard

Table 1: *Acoustic properties observed in normal N, Lombard L, and extrapolated Lombard L-E voices.*

Voice	speech (secs.)	pauses (secs.)	F0 (Hz)	spectral tilt (dB/octave)
Natural speech				
Normal	2.06	-	107.1	-2.02
Lombard	2.32	-	136.8	-1.73
Text-to-speech				
N	2.11	0.16	104.5	-2.09
L	2.80	0.19	145.0	-1.59
L-E	3.05	0.20	144.8	-1.50

(L) voices. We can see that, as expected, the Lombard voice produces sentences with longer duration and longer pauses, greatly increased F0 mean and flattening of the spectral tilt. The spectral tilt reflects changes in both spectral envelope and excitation signal. These values are similar to those of natural Lombard speech.

We mixed the synthetic voices with two noises: speech-modulated noise (ICRA) and speech from a single competing female talker. For intelligibility testing, it is important to avoid floor or ceiling effects on word error rate. Therefore, in order to obtain intelligibility scores in similar ranges for each noise, we mixed them at differing SNRs: -4 dB for speech-modulated noise and -14 dB for the competing talker. Across the different voices we made sure that the root mean square value was the same.

For the listening test we used 32 native English speakers listening to the noisy samples over headphones in soundproof booths and typing in what they heard. Each participant heard six different sentences per condition, i.e., voice and noise type, and each sentence could only be played once. We used the first ten sets of the Harvard sentences (Har, 1969); another one of the sets was used as a practice session which listeners completed before the test proper.

Figs. 3 and 4 show the mean word accuracy rate (WAR) obtained by each voice when mixed with speech-modulated noise and a competing talker respectively, along with 95 % confidence intervals. Figure 3 shows that the Lombard voices L (63.5 %) and L-E (68.1 %) performed better than the normal N voices (40.9 %). The extrapolated voice L-E has a better score than the voice L, however this difference was not statistically significant.

The results obtained for the competing talker situation are displayed in Fig. 4 and show a slightly different trend. All Lombard voices performed significantly better than the normal N voice (36.6 %), in particular the L voice (62.2 %). The extrapolated version, L-E (60.5 %) does not appear to increase intelligibility.

For more details, refer to the following paper:

> Valentini-Botinhao, C., Yamagishi, J. and King, S. (2012). "Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise". Submitted to Interspeech, Portland, USA.

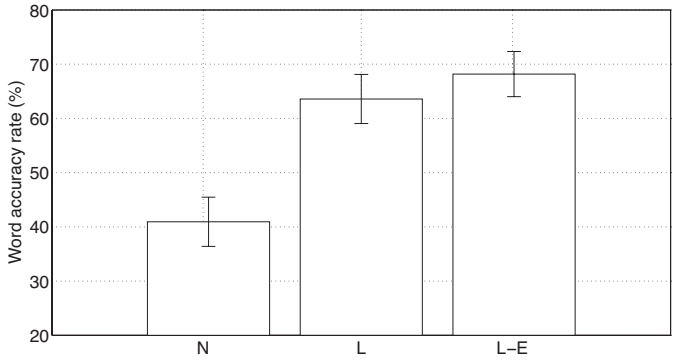

Figure 3: *Word accuracy rates for speech-modulated noise.*

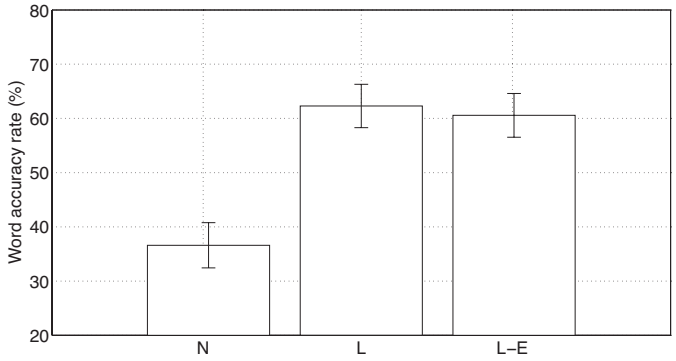

Figure 4: *Word accuracy rates for competing talker.*

2.2 Experiment on the Roger corpus

We conducted a similar experiment using another British male speaker "Roger". The Lombard dataset was recorded while the speaker listened to speech-modulated noise based on multiple speakers (Dreschler et al., 2001) played over headphones at a absolute value of 90 dBA.

Normal voice N was created from a high quality average voice model adapted to 25 TIMIT sentences of the normal speech database. The speaker adaptation algorithm used was CSMAPLR+MAP. Lombard voice L was further adapted using 25 TIMIT sentences from the Lombard speech dataset. We mixed the synthetic voices with ICRA speech-modulated noise at -5 dB, 0 dB and 5 dB SNRs. For the listening test we used 40 native English speakers listening to the noisy samples over headphones in soundproof booths and typing in what he or she heard. We used 78 Matrix sentences.

Table 2 provides an acoustic analysis of the voices – average duration of speech in relative, average spectral tilt, and F0 – across all sentences used in the listening test for the normal (N) and lombard (L) voices. We can see that these values are similar to those of natural Lombard speech, to some extent.

Table 3 shows the mean word accuracy rate (WAR) obtained by each voice when mixed with speech-modulated noise at each SNR. Contrary to the previous experiment on the Nick corpus, we do not see any significant improvement for the Lombard voices at any SNR. In WP5 and WP1, it was also found that this speaker's *natural* Lombard speech does not have higher intelligibility than his normal speech. This implies that mimicking Lombard speech is not always the best strategy for improving the intelligibility.

Table 2: *Acoustic properties observed in normal N and lombard L voices.*

Voice	Relative duration	F0 (Hz)	spectral tilt (dB/octave)
Natural speech			
Normal	1.00	127.8	-2.37
Lombard	1.18	214.6	-1.73
Text-to-speech			
N	1.00	123.3	-2.09
L	0.99	207.2	-1.81

Table 3: *Word accuracy rates (%) for speech-modulated noise.*

Voice	-5 dB	0 dB	5 dB
Text-to-speech			
N	74.7	88.2	94.3
L	77.0	88.1	92.7

3 Task T3.2 Spectral adaptation

In order to perform HMM transformations without requiring speech data from a closely-matched listening situation or without requiring the ideal Lombard speech (rather than just "shouting"), the transforms applied to the HMMs must be treated in a more sophisticated way.

The modifications that human talkers use, identified in WP1 as candidates for study, should be a useful part of such transforms. However, the parameters describing the candidate modifications are typically different from the spectral parameters used for HMM-based speech synthesis systems. In this section, we represent the candidate modification parameters at time t as $\boldsymbol{y}_t$ and describe how we can utilise $\boldsymbol{y}_t$ to control HMM parameters without using additional speech data. Examples of the candidate modification parameters $\boldsymbol{y}_t$ include: formant centre frequencies, articulator movements, and noise spectra.

3.1 Spectral control using the candidate modification vectors

In order to utilise $\boldsymbol{y}_t$ to control HMM parameters (without using additional speech data), we introduce a linear auxiliary (projection) function having a low-dimensional $\boldsymbol{y}_t$ into the output probabilities of HMMs. This is called a multiple-regression HMM (Miyanaga et al., 2004).

The conditional likelihood of a multiple-regression HMM λ, given a candidate modification vector

sequence $\boldsymbol{Y}$, can be written as

$$P(\boldsymbol{X}\,|\lambda,\,\boldsymbol{Y}) = \sum_{\boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{x}_t|\boldsymbol{y}_t), \tag{26}$$

$$b_i(\boldsymbol{x}_t|\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{x}_t;\,\boldsymbol{\mu}_i,\,\boldsymbol{\Sigma}_i), \tag{27}$$

$$= \mathcal{N}(\boldsymbol{x}_t;\,\boldsymbol{H}_i\boldsymbol{\xi}_t + \boldsymbol{z}_i, \boldsymbol{\Sigma}_i), \tag{28}$$

$$\boldsymbol{Y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, ..., \boldsymbol{y}_T^\top]^\top \tag{29}$$

$$\boldsymbol{\xi}_t = \left[\boldsymbol{y}_t^\top, 1\right]^\top \tag{30}$$

where $\boldsymbol{\xi}_t$ is the candidate modification vector vector; $\boldsymbol{H}_i$ is the linear projection matrix for state i. As can be seen from the equation above, the mean vector $\boldsymbol{\mu}_i$ of each state-output distribution consists of the linear projection of the candidate modification parameters $\boldsymbol{y}_t$ and a candidate-modification-independent parameter $\boldsymbol{z}_j$ (and hence it is possible to change voice characteristics by changing $\boldsymbol{y}_t$).

This multiple-regression HMM was initially proposed to improve the accuracy of acoustic modelling for ASR by utilising auxiliary features that are correlated with the acoustic features (Fujinaga et al., 2001); auxiliary features that have been used in ASR include the fundamental frequency (Fujinaga et al., 2001). Auxiliary features that have been often used in TTS are meta-level descriptions of speech such as specific voice characteristics, speaking styles, and emotions so that we can directly manipulate such expressivity, brightness, and emotions of specific words or phrases straightforwardly at the synthesis stage.

In this experiment, we use the candidate modification vectors obtained from WP2's *context-indicators* or from noise itself as the auxiliary features $\boldsymbol{y}_t$ of the multiple-regression HMM.

3.2 Spectral control using feature-space-switched multiple regression HMM

In general, we use more than one linear projection matrix $\boldsymbol{H}_i$, each applying to a subset of the model parameters (known as a "regression class") so the overall effect is a piecewise linear approximation to a globally non-linear transform. Typically, in previous work, several "hard" regression classes are formed by clustering similar model parameters (e.g., by simply using an existing parameter tying tree).

In order to reflect the candidate modification vectors, which are different from the model parameters, we need a better way to form regression classes. First, the regression classes should be formed using the candidate modification vectors since they are different from the model parameters. Second, the regression classes should be "soft" since the candidate modification vectors may be continuous values.

For this purpose, we proposed a new way to form the regression classes in the feature space of the candidate modification vectors (Ling et al., 2011) and have applied it to the multiple-regression HMM, which we call a *feature-space-switched multiple regression HMM*.

In this new proposed method, a GMM model $\lambda^{(G)}$ containing M mixture components is trained in advance using only the candidate modification vectors. Then by assuming that the M clusters correspond to the regression classes, we estimate linear projection matrices of the multiple regression HMM for each of the clusters. Further by using the posterior probabilities of the GMM, we compute a weighted sum of the output probabilities of the HMM, which is equivalent to "soft" regression classes.

We simply rewrite (28) as

$$b_i(\boldsymbol{x}_t|\boldsymbol{y}_t) = \sum_{m=1}^{M} \zeta_m(t)\,\mathcal{N}(\boldsymbol{x}_t;\ \boldsymbol{H}_m\boldsymbol{\xi}_t + \boldsymbol{z}_i, \boldsymbol{\Sigma}_i) \tag{31}$$

$$\zeta_m(t) = P(m|\boldsymbol{y}_t, \lambda^{(G)}) \tag{32}$$

where $\boldsymbol{H}_m$ is a linear projection matrix estimated for the k-th mixture of the GMM $\lambda^{(G)}$ and $\zeta_m(t)$ is the posteriori probability of the candidate modification parameters $\boldsymbol{y}_t$ for the m-th mixture of the GMM $\lambda^{(G)}$.

For more details, please refer to the following paper:

> Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi (2011). "Feature-space transform tying in unified acoustic-articulatory modelling of articulatory control of HMM-based speech synthesis." In Proc. Interspeech, pages 117-120, Florence, Italy, August 2011
>
> Zhenhua Ling, Korin Richmond and Junichi Yamagishi (expected 2012), "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression". Submitted to IEEE Transactions on Audio, Speech, and Language Processing.

Using the proposed new model and the new way of forming the regression classes, we carried out several perceptual experiments to verify that the candidate modification vectors $\boldsymbol{y}_t$, which could be obtained from WP2's *context-indicators*, can control synthetic speech as intended. As we saw in deliverable D2.1, the candidate modification vectors $\boldsymbol{y}_t$ may vary significantly in their type. Therefore we chose several different types of candidate modification vectors for our experiment. Specifically we chose: formants, articulatory movements and noise spectra as the $\boldsymbol{y}_t$ vector and conducted three corresponding experiments.

3.3 Experiment 1 – Formant control

In the first experiment, we used the formant centre frequencies as the candidate modification parameters $\boldsymbol{y}_t$ and carried out vowel and consonant identity perception tests. A phoneme identity perception test is an easy way to assess the controllability of the proposed method.

In this experiment, the proposed model was trained on a database of 1200 sentences recorded by a male, native British English talker. A categorical perception design was used for the perception study. Three sets of speech continua were created. In the first experiment, listeners heard the three-way consonant contrast /bɛt/ ("bet"), /dɛt/ ("det") and /gɛt/ ("get"). In the second experiment, listeners heard the three-way vowel contrast /bit/ ("bit"), /bɛt/ ("bet"), and /bæt/ ("bat"). In all three sets of stimuli, the first three formants were manipulated. All listeners were native speakers of English (28 female, 12 male) aged between 18 years and 38 years (average age: 23 years).

Figure 5 shows the distribution of listener responses to the two sets of vowel stimuli. The three end/midpoint stimuli received nearly 100% correct identification (token 'e2a-3', the /bæt/ endpoint, did in fact receive 100% correct responses). Additionally, the boundaries between the three phone categories are very steep, with little overlap.

Figure 6 shows the distributions of listener responses to the consonant contrast continuum. The first point to observe is that none of the end-/midpoint consonant stimuli ('d', 'd2b-4', and 'd2g+4';

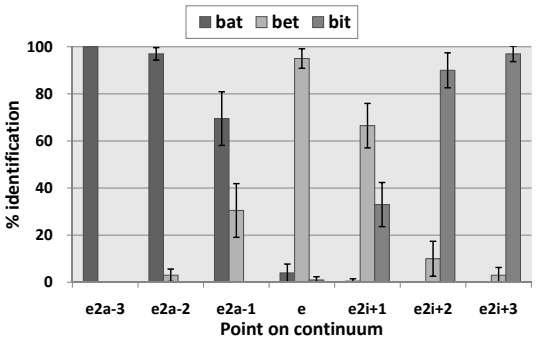

Figure 5: *Average listener responses to the vowel continuum produced by the* GMM-HMM *system (top graph), and by the* DT-HMM *system (bottom graph). In both graphs, the original, unmodified /dɛt/ token is marked as 'e'. The same token modified to contain vowel formant steady-state values appropriate for /bæt/ is marked as 'e2a-3', while the 'e' token modified to contain vowel formant steady-state values appropriate for /bit/ is marked as 'e2i+3'. Intermediate stimuli change in* $+/-50$ *Hz steps between these three points (see text for details). Error bars represent 95% confidence intervals.*

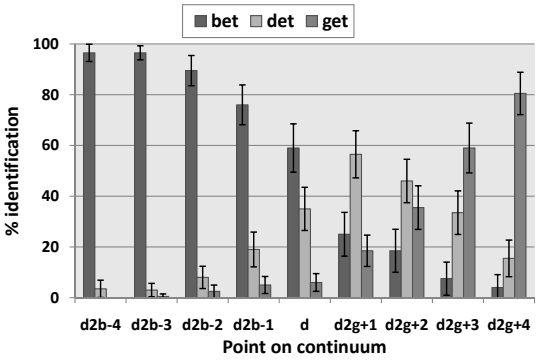

Figure 6: *Average listener responses to stimuli along the consonant contrast continuum. The original, unmodified /dɛt/ token is marked as 'd'. The same token resynthesised using the vowel formant onset contour from /bɛt/ is marked 'd2b-4'; 'd2g+4' indicates the same token resynthesised using the formant onset contour from /gɛt/. Intermediate stimuli change in equal steps between these three points (see text for details). Error bars represent 95% confidence intervals.*

see reference for details) received 100% "det", "bet" and "get" responses. Similarly, there is a certain amount of overlap at category boundaries. However, it is also clear that listeners were able to identify the stimuli as being from three different phone categories. These results imply that the proposed method has an ability to control formants of synthetic speech precisely and thus it may be used for e.g. enhancing vowel space, which is a typical phenomenon observed in Lombard speech data.

For more details, please refer to the following papers:

Catherine Mayo, Ming Lei, Junichi Yamagishi, Korin Richmond and Zhen-Hua Ling (2012), "A perceptual study of formant-controlled HMM-based speech synthesis." Submitted to Interspeech, Portland, USA.

Ming Lei, Junichi Yamagishi, Korin Richmond, Zhen-Hua Ling, Simon King, and Li-Rong Dai (2011). "Formant-controlled HMM-based speech synthesis." In Proc. Interspeech, pages 2777-2780, Florence, Italy, August 2011

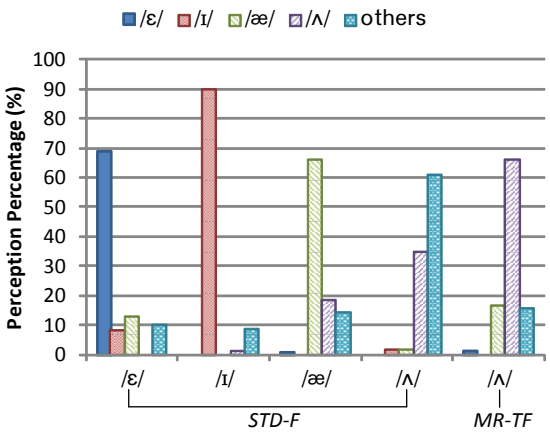

Figure 7: Vowel identity perception results for synthesising different vowels using the *STD-F* system and creating vowel /ʌ/ by articulatory control using the *MR-TF* system.

3.4 Experiment 2 – Articulatory control

In the second experiment, we used articulator positions as the candidate modification parameters $\boldsymbol{y}_t$ and carried out a vowel identity perception test. In this experiment, we created a new target /ʌ/ vowel by only manipulating the articulators via a model that was trained on speech data that did not include any tokens of /ʌ/. In other words, the aim of this experiment is to confirm that speech transformation can be achieved without requiring speech data from a closely-matched condition.

These results are shown in Fig. 7. We see that only 35% of the synthesised vowels /ʌ/ were perceived correctly using the standard system notated *STD-F*, due to the lack of acoustic training samples for this vowel. Using the proposed system notated *MR-TF* and manipulating the articulatory features properly, this percentage increased to 66.25%, which is close to the perception accuracy of synthetic vowels /ɛ/ (68.75%) and /æ/ (66.25%) using the *STD-F* system.

For more details, please refer to the following paper:

> Zhen-Hua Ling, Korin Richmond and Junichi Yamagishi (2012), "Vowel Creation by Articulatory Control in HMM-based Parametric Speech Synthesis." Submitted to Interspeech, Portland, USA.

3.5 Experiment 3 – Noise-based control

In the third experiment, which is still in progress, we use the noise spectrum directly as the candidate modification parameter $\boldsymbol{y}_t$. This may be viewed as one possible way to integrate *context-indicators* into HMM-based speech synthesis, which is planned as a Year 3 task (T3.6).

To train the model, we need a parallel corpus. In the preceding experiments that was either speech + formant tracks, or speech + articulator positions (recorded using EMA). Now the parallel corpus consists of speech and the noise that the speaker was listening to over headphones. To conduct this experiment, we created a new corpus comprising parallel recordings of speech, articulator positions and the noise signal. The speech was uttered by a professional speaker in a variety of noises – see Figure 8.

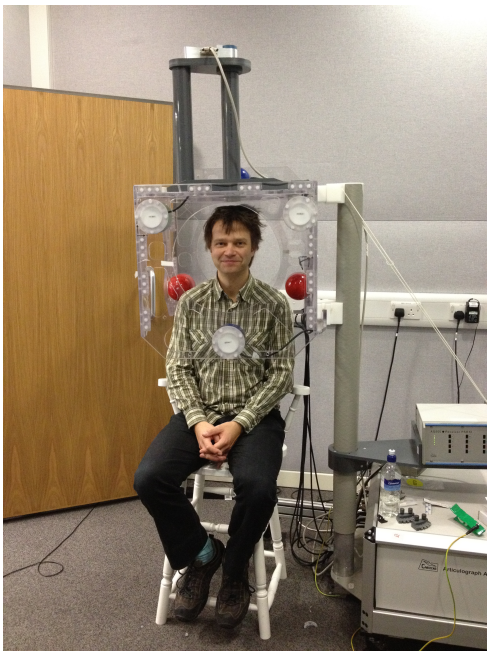

Figure 8: Electromagnetic articulography (EMA) uses a special device to directly track and record the movement of sensor coils attached to the tongue or other articulators by using an electromagnetic field. In addition to articulatory movements, we recorded the speech signal and the noise that the speaker was listening to. Video was also recorded. EMA is much cheaper, safer and quieter (essentially silent equipment) than X-ray microbeam methods.

For more details of this corpora, refer to D5.1. Several experiments using this corpora are already planned and they are expected to be reported in the deliverables due at the end of year 3.

4 Task T3.3 Temporal/prosodic adaptation

Task T3.2 concerned spectral adaptation, and we now examine the separate effects of temporal/prosodic adaptation.

4.1 Direct use of speaker adaptation for temporal and prosodic adaptation

As we described in the background material at the start of this deliverable, speaker adaptation can be applied to temporal and prosodic features in the same way as to the spectral features (Yamagishi and Kobayashi, 2007). We first assessed the intelligibility improvements due to conventional temporal and prosodic speaker adaptation compared to spectral adaptation. The experimental conditions are identical to ones mentioned in Sect. 2.1.

In this evaluation, in addition to the normal voice N and Lombard voice L used in the previous experiment, Voice N-L was also created from voice N but this time only the Mel cepstral coefficients were adapted to the Lombard data. In other words, N-L only uses spectral adaptation and no temporal/prosodic adaptation. The goal of this experiment is to separate out and quantify the effect on intelligibility of

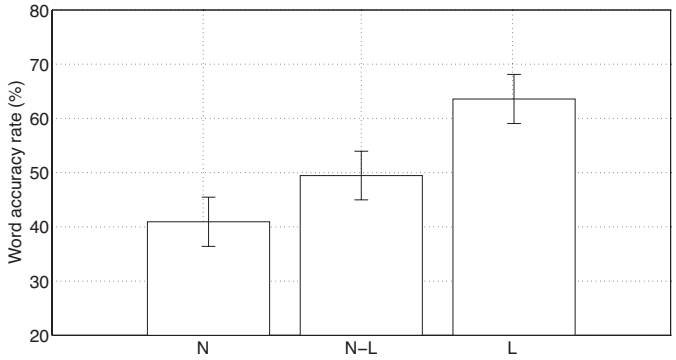

Figure 9: *Word accuracy rates for speech-modulated noise.*

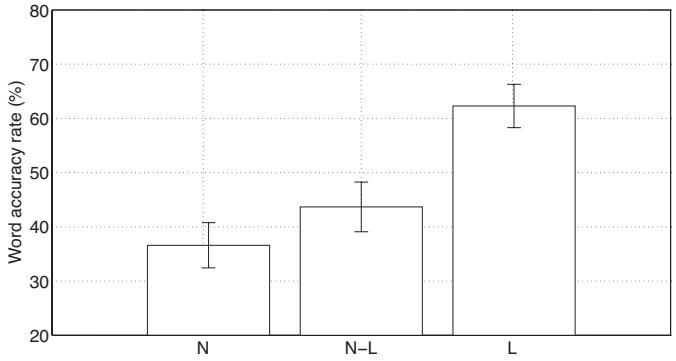

Figure 10: *Word accuracy rates for competing talker.*

temporal/prosodic adaptation.

Figs. 9 and 10 show the mean word accuracy rate (WAR) obtained by each voice when mixed with speech-modulated noise and a competing talker respectively, along with $95\,\%$ confidence intervals. Fig. 9 shows that voice N-L ($49.4\,\%$) is significantly better than voice N ($40.9\,\%$). The Lombard voice L ($63.5\,\%$), where prosodic and temporal features are adapted as well as the spectral features, is significantly better than the N-L voice. The intelligibility gains obtained by the full Lombard voice L over the N-L voice reflect the impact of changes in duration patterns, F0 and the aperiodicity parameters that define the excitation signal. We can see, then, that there is a lot to gain from modifying those parameters in addition to the spectral ones.

The results obtained for the competing talker situation are displayed in Fig. 10 and show a slightly different trend. The difference between the normal voice N ($36.6\,\%$) and the N-L voice ($43.6\,\%$) was not statistically signifiant, whereas the L voice ($62.2\,\%$) was found to be significantly better than the others. For the competing talker situation, spectral changes seem to contribute less than in speech-modulated noise. For the competing talker, duration stretches as well as F0 increases are more important.

4.2 Analysis of F0 and duration modification

The analysis of temporal and prosodic adaptation above revealed that duration stretches as well as F0 increases are important for intelligibility improvement. However, it does not show which one contributes

most. Next, temporal and prosodic adaptation are separately evaluated in order to obtain further insight.

We applied the following individual simple modifications to the synthesized material:

- changes in the fundamental frequency: low / high
- changes in the speaking rate (Yoshimura et al., 1999): slow / fast

and performed listening tests to measure the intelligibility of synthetic speech in various noises, using 88 native English listeners.

From a detailed analysis of the listening tests, we found that a) changing the speaking rate contributes intelligibility improvements in some conditions and that b) increasing fundamental frequency did not seem to provide any significant gains in intelligibility, although increased F0 is observed in natural Lombard speech. Our results are consistent with another study, in which natural speech was modified, reported by Lu and Cooke (2009).

For more details, refer to the following paper:

> Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King, (2011) "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" Proc. Interspeech 2011, Florence, Italy, August 2011

4.3 Improved temporal modelling

Based on these analysis results, one of the goals in the next year is to look for more robust and flexible temporal modeling and modification methods to improve the intelligibility of speech in noise. Towards this goal, we have explored improved temporal modelling schemes at a more fundamental level, developing alternatives to discrete-state hidden Markov models. Two new approaches for temporal and trajectory modeling have been proposed within LISTA. These methods are still under development.

The first approach is called the "intermediate-state HMM", where the current state variable in a left-to-right HMM can be a real number in-between the integer states. This has be shown to enable both natural transitions and durations with a small set of parameters tied to integer states only. Intermediate-state models can be trained efficiently using the EM algorithm. A further advanced temporal trajectory model called the "Gaussian process dynamical model" has also been proposed and is under investigation.

For more details, refer to the following papers:

> Gustav Eje Henter, W. Bastiaan Kleijn "Intermediate-state HMMs to capture continuously-changing signal features" Proc. Interspeech 2011 Florence, Italy, August 2011
>
> Gustav Eje Henter, Marcus R. Frean, W. Bastiaan Kleijn "Gaussian process dynamical models for nonparametric speech representation and synthesis", Proc. ICASSP 2012, Kyoto Japan, March 2012.

4.4 Other prosodic modifications developed in WP1 and WP4

In addition to the modifications and approaches above, several prosodic modifications for recorded speech were proposed in other WPs in Year 2. They may be combined with text-to-speech levels and this

is planned for evaluation in Year 3. An example of prosodic modification is the rhythmogram, which is a technique to extract the rhythm pattern of speech – this may be used as the candidate modification parameter for controlling the duration probabilities in HMM-based speech synthesis. Please refer to D4.2 and the following publication:

> Julian Villegas, Martin Cooke (2012), "Maximising objective speech intelligibility by local F0 modulation", Submitted to Interspeech 2012, Portland, USA.

5 Task T3.4 Applying speech modifications in the vocoder domain

In addition to model-level modifications, it is also possible to manipulate the input parameters to the vocoder to modify the output waveform. This is a task to be completed in Year 3; here we present our progress so far.

5.1 LSP vocoder – LSP shift

In the first experiment for the vocoder domain modifications, we adopted "frequency shift" of Line Spectral Pairs (LSPs) (McLoughlin and Chance, 1997). This is a very simple operation in the LSP and MGC-LSP vocoder domain, and we found that this is a very effective approach to improve intelligibility. This is because when the LSPs are shifted towards higher frequencies, the spectral tilt becomes flatter and the energy is reallocated into higher frequency regions.

We conducted a perceptual evaluation using several types of noises at various SNRs. The largest improvements on average word accuracy were in the presence of car noise. For the lowest SNR case there was an improvement of word accuracy 13% to 61% and for higher SNRs the word accuracy improved from 38% to 72% and from 42% to 80%. However, shifting the LSPs does not always increase intelligibility. For high frequency noise, a large shift in the LSPs results in a significant drop in word accuracy, while small shifts give significant improvements.

This result suggests that the optimal value of modification strength depends on the noise. Some attempts to guide the LSP shift using ASR-based intelligibility objective measures (reported in D2.2) were also made in year 2.

For more details, refer to the following paper:

> C. Valentini-Botinhao, J. Yamagishi, and S. King (2011) "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?", in Proc. Interspeech, Florence, Italy, August 2011

5.2 Mel-cepstrum vocoder – Cepstral analysis based on the Glimpse proportion measure

Objective intelligibility measures may be used for directly manipulating the vocoder parameters such as cepstra so that the resulting synthetic speech has better intelligibility. We have proposed cepstral and mel-cepstral modification methods based on the Glimpse proportion measure (Cooke, 2006) for intelligibility

enhancement of speech in noise. The Glimpse proportion measure was found to be well correlated with intelligibility of HMM-based synthetic speech in noise. The proposed cepstral analysis method modifies the cepstral coefficients of clean speech to enhance intelligibility as given by the Glimpse proportion measure for speech intelligibility in noise.

To evaluate the method we built eight different voices from normal read-text speech data from a male speaker. Some voices were also built from Lombard speech data produced by the same speaker. Listening experiments with speech-modulated noise and with a single competing talker indicate that our method significantly improves intelligibility when compared to unmodified synthetic speech.

Whilst voices built from actual Lombard speech outperformed the proposed method (particularly for the competing talker case), compared to a voice using only the spectral parameters from Lombard speech, the proposed method obtains comparable or higher performance. However, this proposed method does not require recorded Lombard speech and so is easy to apply to other speakers, noises and SNRs.

For more details, refer to the following papers:

Valentini-Botinhao C., Yamagishi, J., and King, S. (2011) "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?", in Proc. Interspeech, Florence, Italy, August 2011

Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S. and Zen, H. (2012), "Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise". In Proc. ICASSP, Kyoto, Japan

Valentini-Botinhao, C., Yamagishi, J. and King, S. (2012), "Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise". Submitted to Interspeech, Portland, USA.

5.3 Harmonic model vocoder

We have also proposed some techniques for manipulating harmonic model parameters to increase speech intelligibility in noise, without modifying the energy of the signal. They involve a two-step transformation. During the first step, the spectral slope is increased to mimic the effect of higher vocal effort. During the second step, the energy of the signal is redistributed over time to amplify meaningful low-energy parts of the signal. This transformation operates on the harmonic amplitudes only and can be easily integrated into a text-to-speech synthesizer without increasing the computational costs.

For more details, refer to the following paper:

Daniel Erro, Yannis Stylianou, Eva Navas and Inma Hernaez, (2012) "Implementation of simple spectral techniques to enhance the intelligibility of speech using a harmonic model", Submitted to Interspeech 2012

6 Task T3.5 Applying waveform processing techniques to synthetic speech

Various modifications techniques for recorded speech can be applied as a simple post processing of synthetic speech waveforms. This is a task to be completed during 3. The modifications for recorded speech have been developed in other WPs, including:

Tudor-Catalin Zorila, Varvara Kandia, Yannis Stylianou, (2012) “Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression”, Submitted to Interspeech 2012

Petko N. Petkov, W. Bastiaan Kleijn, Gustav Eje Henter, (2012) “Enhancing subjective speech intelligibility using a statistical model of speech”, Submitted to Interspeech 2012, Portland, USA.

Yan Tang, Martin Cooke (2012) “Optimised spectral weightings for noise-dependent speech intelligibility enhancement”, Submitted to Interspeech 2012, Portland, USA.

References

(1969). IEEE recommended pratice for speech quality measurements. *Audio and Electroacoustics, IEEE Transactions on*, 17(3):225 – 246.

Chien, J., Wang, H., and Lee, C. (1997). Improved Bayesian learning of hidden Markov models for speaker adaptation. In *Proc. ICASSP-97*, pages 1027–1030.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 119(3):1562–1573.

Digalakis, V. and Neumeyer, L. (1996). Speaker adaptation using combined transformation and Bayesian methods. *IEEE Trans. Speech Audio Process.*, 4:294–300.

Dreschler, W., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. *Audiology*, 40(3):148–57.

Fujinaga, K., Nakai, M., Shimodaira, H., and Sagayama, S. (2001). Multiple-regression hidden Markov model. In *ICASSP*, pages 513–516.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. ICASSP*, pages 137–140.

Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language,*, 12 (2):75–98.

Gauvain, J. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. Speech Audio Processing*, 2(2):291–298.

Kim, D. Y., Umesh, S., Gales, M. J. F., Hain, T., and Woodland, P. C. (2004). Using VTLN for broadcast news transcription. In *Proc. of ICSLP*, pages 1953–1956, South Korea.

Lee, L. and Rose, R. (1998). A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing,*, 6:49–60.

Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang.*, 9:171–185.

Ling, Z.-H., Richmond, K., and Yamagishi, J. (2011). Feature-space transform tying in unified acoustic-articulatory modelling for articulatory control of HMM-based speech synthesis. In *Interspeech*, pages 117–120.

Lu, Y. and Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Comm.*, 51(12):1253–1262.

McDonough, J. W. (2000). *Speaker Compensation with All-Pass Transforms*. PhD thesis, John Hopkins University.

McLoughlin, I. and Chance, R. (1997). LSP-based speech modification for intelligibility enhancement. In *Proc. Digital Signal Processing*, volume 2, pages 591–594, Santorini , Greece.

Miyanaga, K., Masuko, T., and Kobayashi, T. (2004). A style control technique for HMM-based speech synthesis. In *Proc. Interspeech*, pages 1437–1439.

Odell, J. (1995). *The use of context in large vocabulary speech recognition*. PhD thesis, Cambridge Univ., Cambridge, U.K.

Pitz, M. and Ney, H. (2005). Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing,*, 13:930–944.

Shinoda, K. and Lee, C. (2001). A structural Bayes approach to speaker adaptation. *IEEE Transactions on Speech Audio Processing*, 9:276–287.

Shiohan, O., Myrvoll, T., and Lee, C. (2002). Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer, Speech and Language*, 16(3):5–24.

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994). Mel-generalized cepstral analysis – A unified approach to speech spectral estimation. In *Proc. of ICSLP*, volume 3, pages 1043–1046.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (2002). Multi-space probability distribution HMM. *IEICE Trans. Inf. Syst.*, E85-D(3):455–464.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pages 1315–1318.

Woodland, P. C. (2001). Speaker adaptation for continuous density HMMs: A review. In *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition*, pages 11–19.

Yamagishi, J. and Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans. Inf. & Syst.*, E90-D(2):533–543.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Speech, Audio & Language Process.*, 17(1):66–83.

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*, pages 2347–2350.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speaker interpolation for HMM-based speech synthesis system. *Acoustical Science and Technology*, 21(4):199–206.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2004). Hidden semi-Markov model based speech synthesis. In *Proc. of the International Conference on Spoken Language Processing*, pages 1397–1400, Korea.