DECISION TREE-BASED CONTEXT CLUSTERING BASED ON CROSS VALIDATION AND HIERARCHICAL PRIORS

Heiga Zen and M. J. F. Gales

Toshiba Research Europe Ltd. Cambridge Research Laboratory, Cambridge, United Kingdom
heiga.zen@crl.toshiba.co.uk

ABSTRACT

The standard, ad-hoc stopping criteria used in decision tree-based context clustering are known to be sub-optimal and require parameters to be tuned. This paper proposes a new approach for decision tree-based context clustering based on cross validation and hierarchical priors. Combination of cross validation and hierarchical priors within decision tree-based context clustering offers better model selection and more robust parameter estimation than conventional approaches, with no tuning parameters. Experimental results on HMM-based speech synthesis show that the proposed approach achieved significant improvements in naturalness of synthesized speech over the conventional approaches.

Index Terms— decision tree-based context clustering, cross validation, hierarchical priors, HMM-based speech synthesis

1. INTRODUCTION

In both speech recognition and synthesis, context-dependent hidden Markov models (HMMs) are widely used. Speech recognition systems typically use triphone or quinphone HMMs. In speech synthesis, segmental, prosodic, and linguistic contexts in addition to phonetic contexts (sometimes called "fullcontexts") are often used (e.g. [1]). There are a huge number of possible combinations of contexts; thus it is almost impossible to cover all possible combinations of contexts with a finite set of training data. To address this problem, decision tree-based context clustering was proposed [2]. This is a top-down, data-driven clustering technique based on a greedy tree growing algorithm. It clusters similar context-dependent HMM states (or streams [3]) into the same class by building decision trees so as to maximize (or minimize) an objective function using predefined split candidates (questions about contexts). Model parameters are then tied across context-dependent HMM states associated with the same class (terminal node). Typically, the maximum likelihood (ML) criterion is used to select the best split candidate. Decision tree-based context clustering yields a compact set of unique models which can represent every possible combination of contexts.

Although decision tree-based context clustering based on the ML criterion works reasonably well, there are two problems.

1. Ad-hoc stopping criterion; the likelihood of the model given the training data monotonically increases as the number of terminal nodes increases. Without using any ad-hoc stopping criteria such as thresholds for increase in log likelihoods or minimum occupancy counts, the model becomes almost identical to context-dependent HMMs without clustering.
2. No regularization; parameters estimated by the ML criterion often overfit to the training data if the number of training samples is small.

There have been many attempts to address these problems. One solution for the first problem is using an information criterion such as Akaike's information criterion (AIC), minimum description length (MDL) principle [4], or Bayesian information criterion (BIC). These information criteria have theoretically-derived penalty terms that balance model accuracy and complexity. However, a generic penalty term often does not work well due to the mismatch between the assumptions made in these information criteria and real data. For example, they assume that observations are independent identically distributed (i.i.d.) random variables, but speech observations are dependent on each other. Therefore, in practice, their penalty terms are often empirically scaled [4]. The Bayesian approach [5] also offers an automatic stopping criterion. It uses the lower bound of the log marginal likelihood derived from a variational approximation as the objective function to be maximized. As the Bayesian approach incorporates prior distributions over the model parameters, it can also address the second problem. However, this approach typically uses static, global priors whose hyper-parameters are manually set or determined from statistics at root nodes of decision trees; thus their regularization abilities are limited.

Recently, cross validation (CV) [6] was introduced to decision tree-based context clustering [7, 8]. Cross validation provides estimates of the generalization error and has been used in various stages in machine learning such as model selection and parameter tuning. Decision tree-based context clustering based on CV offers automatic stopping criterion, i.e. if the estimate of the generalization error increases (CV likelihood decreases) with the best split, tree growth is stopped. This has worked successfully in both speech recognition [8] and synthesis [9]. However, it still relies on the ML estimates of parameters. If the number of training samples associated with a node is small, it may not give reliable estimates of the model parameters. As decisions (selecting the best questions and stopping the tree growth) are based on these unreliable estimates of parameters, the resulting tree may not yield robust model parameters.

To address this problem, this paper proposes decision tree-based context clustering based on CV and hierarchical priors. In addition to CV, the proposed approach uses priors to regularize parameter estimation. Unlike the conventional Bayesian clustering approach, dynamic, hierarchical priors similar to structural maximum a posteriori (SMAP) estimation [10] are used. The combination of CV and hierarchical priors within decision tree-based context clustering gives more robust parameter estimation and better model selection. Furthermore, regularization parameters of hierarchical priors can be selected by CV within the clustering process.

The rest of this paper is organized as follows. Section 2 describes the proposed approach for decision tree-based context clustering based on CV and hierarchical priors. Section 3 shows experimental results in HMM-based speech synthesis [3]. Concluding remarks and future plans are given in the final section.

2. DECISION TREE-BASED CONTEXT CLUSTERING BASED ON CV AND HIERARCHICAL PRIORS

Decision tree-based context clustering performs a top-down clustering using questions about contexts. Model parameters (typically mean vectors and covariance matrices) of HMM states (or streams) clustered to the same class (terminal nodes) are tied. This section describes the proposed approach of decision tree-based context clustering based on CV and hierarchical priors. Please refer to [2] for the details of decision tree-based context clustering technique.

2.1. Cross Validation

The efficient implementation of decision tree-based context clustering based on K-fold CV [8] uses CV log likelihood as its objective function to be maximized. It first divides training data $\mathcal{D}$ into K subsets $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$, where $\mathcal{D} = \bigcup_{i=1}^{K} \mathcal{D}^{(i)}$ and $\mathcal{D}^{(i)} \bigcap \mathcal{D}^{(j)} = \emptyset$ (empty) for any i and j. The occupancy counts and first- and second-order statistics of $\mathcal{D}^{(k)}$ associated with a node s are given as

$$\Gamma_s^{(k)} = \sum_{t \in \mathcal{D}^{(k)}} \sum_{m \in \boldsymbol{M}_s} \gamma_m(t), \tag{1}$$

$$\boldsymbol{\nu}_s^{(k)} = \sum_{t \in \mathcal{D}^{(k)}} \sum_{m \in \boldsymbol{M}_s} \gamma_m(t)\, \boldsymbol{o}_t, \tag{2}$$

$$\boldsymbol{\Omega}_s^{(k)} = \sum_{t \in \mathcal{D}^{(k)}} \sum_{m \in \boldsymbol{M}_s} \gamma_m(t)\, \boldsymbol{o}_t \boldsymbol{o}_t^\top, \tag{3}$$

where $\Gamma_s^{(k)}$, $\boldsymbol{\nu}_s^{(k)}$, and $\boldsymbol{\Omega}_s^{(k)}$ correspond to the occupancy counts, first- and second-order statistics, $\boldsymbol{M}_s$ denotes a set of context-dependent HMM states (or streams) associated with s, and $\gamma_m(t)$ is the posterior probability of a context-dependent HMM state m given the observation vector at frame t, $\boldsymbol{o}_t$. While evaluating the k-th fold, the subsets $\mathcal{D}^{(\bar{k})} = \bigcup_{i \neq k} \mathcal{D}^{(i)}$ are used to estimate the model parameters, where $\bar{k} = \{1, \ldots, k-1, k+1, \ldots, K\}$. The ML estimates of the mean vector $\boldsymbol{\mu}_s^{(k)}$ and the covariance matrix $\boldsymbol{\Sigma}_s^{(k)}$ associated with s estimated from $\mathcal{D}^{(\bar{k})}$ are given as

$$\boldsymbol{\mu}_s^{(k)} = \frac{\boldsymbol{\nu}_s^{(\bar{k})}}{\Gamma_s^{(\bar{k})}}, \qquad \boldsymbol{\Sigma}_s^{(k)} = \frac{\boldsymbol{\Omega}_s^{(\bar{k})}}{\Gamma_s^{(\bar{k})}} - \boldsymbol{\mu}_s^{(k)} \boldsymbol{\mu}_s^{(k)^\top}, \tag{4}$$

where

$$\Gamma_s^{(\bar{k})} = \sum_{i \neq k} \Gamma_s^{(i)}, \quad \boldsymbol{\nu}_s^{(\bar{k})} = \sum_{i \neq k} \boldsymbol{\nu}_s^{(i)}, \quad \boldsymbol{\Omega}_s^{(\bar{k})} = \sum_{i \neq k} \boldsymbol{\Omega}_s^{(i)}, \tag{5}$$

Then the remaining single subset $\mathcal{D}^{(k)}$ is used to evaluate the log likelihood of the estimated model. The log likelihood of the model associated with s estimated by the training subsets $\mathcal{D}^{(\bar{k})}$ given the evaluation subset $\mathcal{D}^{(k)}$ is calculated as

$$\begin{aligned}\mathcal{L}(\lambda_s^{(k)}; \mathcal{D}^{(k)}) = \frac{1}{2} \Big\{ \Gamma_s^{(k)} \log\left(2\pi \left|\boldsymbol{\Sigma}_s^{(k)}\right|\right) + \mathrm{tr}\left(\boldsymbol{\Omega}_s^{(k)} \boldsymbol{\Sigma}_s^{(k)^{-1}}\right) \\ - 2\boldsymbol{\mu}_s^{(k)^\top} \boldsymbol{\Sigma}_s^{(k)^{-1}} \boldsymbol{\nu}_s^{(k)} + \Gamma_s^{(k)} \boldsymbol{\mu}_s^{(k)^\top} \boldsymbol{\Sigma}_s^{(k)^{-1}} \boldsymbol{\mu}_s^{(k)} \Big\},\end{aligned} \tag{6}$$

where $\lambda_s^{(k)} = \left\{ \boldsymbol{\mu}_s^{(k)}, \boldsymbol{\Sigma}_s^{(k)} \right\}$. The above process is then repeated over K folds, with each of the K subsets used exactly once as the validation data. Based on this procedure, the CV log likelihood (estimate of the negative generalization error) of the node s is calculated as

$$\mathcal{L}(s; \mathcal{D}) = \sum_{k=1}^{K} \mathcal{L}(\lambda_s^{(k)}; \mathcal{D}^{(k)}). \tag{7}$$

When the node s is split into s_+^q and s_-^q by a question q, the gain in the CV log likelihood is given as

$$\Delta\mathcal{L}(s; \mathcal{D}, q) = \mathcal{L}(s_+^q; \mathcal{D}) + \mathcal{L}(s_-^q; \mathcal{D}) - \mathcal{L}(s; \mathcal{D}). \tag{8}$$

During the clustering process, the best question that maximizes $\Delta\mathcal{L}(s; \mathcal{D}, q)$ is selected for each node as

$$\hat{q} = \arg\max_q \Delta\mathcal{L}(s; \mathcal{D}, q). \tag{9}$$

Since it is based on CV, $\max_q \Delta\mathcal{L}(s; \mathcal{D}, q)$ can take negative values. It indicates that the best split increases the generalization error indicating the model maybe overfits to the training data. The appropriate stopping criterion is to stop when $\max_q \Delta\mathcal{L}(s; \mathcal{D}, q) < 0$.

2.2. Hierarchical Priors

Decision tree-based context clustering based on CV can avoid selecting splits which make the model overfit to the training data. However, it is still based on the ML estimates of parameters. If the number of training samples associated with a node is small, reliable estimates of model parameters may not be obtained. As decisions (selecting the best questions and stopping the tree growth) are based on these unreliable estimates of parameters, the resulting tree may not yield robust model parameters and appropriate model structure. To avoid this problem, this paper introduces hierarchical priors. It is known that the use of priors can regularize parameter estimation and reduce, or eliminate the overfitting problem. Rather than using static, global priors, this paper introduces dynamic, hierarchical priors similar to structural MAP estimation [10]; priors are propagated along with the decision tree structure.

A count-smoothing approach for incorporating dynamic priors [11] is used. Statistics for estimating mean and covariance matrices are based on interpolating statistics of the current and parent nodes, given by

$$\Gamma_s^{(\bar{k})} = \sum_{i \neq k} \Gamma_s^{(i)} + \tau_s, \tag{10}$$

$$\boldsymbol{\nu}_s^{(\bar{k})} = \sum_{i \neq k} \boldsymbol{\nu}_s^{(i)} + \tau_s \frac{\boldsymbol{\nu}_{s^p}^{(\bar{k})}}{\Gamma_{s^p}^{(\bar{k})}}, \tag{11}$$

$$\boldsymbol{\Omega}_s^{(\bar{k})} = \sum_{i \neq k} \boldsymbol{\Omega}_s^{(i)} + \tau_s \frac{\boldsymbol{\Omega}_{s^p}^{(\bar{k})}}{\Gamma_{s^p}^{(\bar{k})}}, \tag{12}$$

where s^p denotes the parent node of s and τ_s is the regularization parameter to scale the prior statistics for s. The prior statistics are normalised so that they effectively contribute τ_s frames to the final statistics. The log likelihood for each fold and the CV log likelihood of s for a given τ_s, $\mathcal{L}(\lambda_s^{(k)}, \tau_s; \mathcal{D}^{(k)})$ and $\mathcal{L}(s, \tau_s; \mathcal{D})$, can be computed in the same way as Eqs. (6) and (7), respectively, based on the interpolated statistics given in Eqs. (10)–(12) instead Eq. (5). Prior statistics for the root node s_0 are set as

$$\Gamma_{s_0^p}^{(\bar{k})} = 1, \quad \boldsymbol{\nu}_{s_0^p}^{(\bar{k})} = \boldsymbol{0}, \quad \boldsymbol{\Omega}_{s_0^p}^{(\bar{k})} = \boldsymbol{I}, \quad k = 1, \ldots, K, \tag{13}$$

where $\boldsymbol{0}$ and $\boldsymbol{I}$ correspond to a zero vector and an identity matrix. As a result, the proposed approach can propagate priors along with the decision tree structure while building decision trees.

2.3. Determination of Regularization Parameter τ_s

Cross validation can also be used to determine values of tuning parameters. As the proposed approach performs CV to obtain the "likelihood" at each split, it can select the best regularization parameter for each split.

For each split, the value of τ_s that maximizes the CV log likelihood can be selected from a set of pre-defined candidate values as

$$\hat{\tau}_s = \arg\max_{\tau_s \in \mathcal{T}} \mathcal{L}(s, \tau_s; \mathcal{D}), \tag{14}$$

where $\mathcal{T}$ denotes the set of pre-defined candidate values of τ_s. This allows the influence of the prior to vary as the quantity and the nature of the data alters.

By setting a large enough value to the number of CV folds K and a set of sufficiently diverse candidate values of τ_s in $\mathcal{T}$, the proposed approach can select the appropriate model structure and give reliable estimates of model parameters with no tuning parameters. The rest of this paper refers to the proposed approach as cross validation structural MAP (CVSMAP).

3. EXPERIMENTS

As HMM-based speech synthesis uses many contexts and single Gaussian component per state, sizes and the structure of decision trees directly affect the quality of synthesized speech. Therefore, the performance of the proposed approach was evaluated in a speech synthesis experiment.

3.1. Experimental conditions

4,624 sentences in US English uttered by a professional female speaker were used for training. The test set consisted of 508 utterances not included in the training data. The sampling frequency was 48 kHz, later down-sampled to 16 kHz. The speech analysis conditions and model topologies of Nitech-HTS 2005 [12] were used. 39-order mel-cepstral coefficients were extracted from the smoothed periodogram. The fundamental frequency (F_0) values of the recordings were automatically extracted using the voting method [13]. No manual correction of the extracted F_0 values was performed. After repeating the decision tree-based context clustering based on the MDL criterion and five runs of embedded reestimation four times, the parameter sharing structure was untied then one embedded reestimation was performed to collect the statistics. Then four clustering algorithms were run using the collected statistics, which were

MDL Decision tree-based context clustering based on the MDL stopping criterion [4].

SMAP Decision tree-based context clustering based on the structural MAP criterion. The proposed hierarchical regularization was used but CV was not. The regularization parameter τ_s was fixed to a global value (0.1, 1, or 10).

CVML Decision tree-based context clustering based on CV [8].

CVSMAP Decision tree-based context clustering based on CV and hierarchical priors. The set of candidate values of τ_s was $\mathcal{T} = \left\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4, 10^5\right\}$.

For all systems no embedded reestimation was performed after the above clustering step. Thus the same statistics were used for building decision trees of all models. Based on the previous work [8, 9], the number of folds for CV was set to $K = 10$. The total numbers of context-dependent models and questions were 176,531 and 3,294, respectively. For synthesis, the speech parameter generation algorithm considering global variance (GV) [14] (context-independent, without silence) was used.

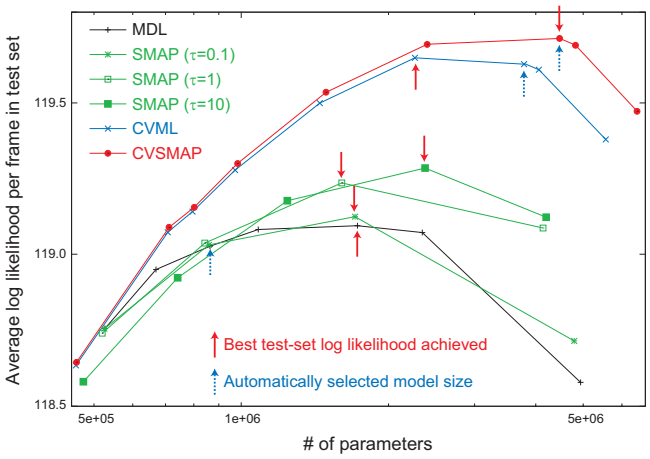

Fig. 1. Average log likelihood per frame of the MDL, SMAP, CVML, and CVSMAP models over the test set.

3.2. Experimental results

Figure 1 plots the average log likelihoods of the models over the test set. The model size automatically selected by the MDL was smaller than CVML and CVSMAP and not optimal in the sense of the test-set log likelihood. Because SMAP included hierarchical regularization, it gave better test-set log likelihoods when the size of model was large. However, CVML and CVSMAP gave better test-set log likelihoods than SMAP. The test-set log likelihoods of CVML and CVSMAP were similar when the size of model was small (the numbers of model parameters $< 10^6$). When the numbers of model parameters were large, CVSMAP gave the better test-set log likelihoods than CVML. This indicates that CVSMAP was more robust than CVML and could avoid overfitting. The model size automatically selected by CVML was slightly overfitted to the training data. On the other hand, the model size automatically selected by CVSMAP gave the best test-set log likelihood.

A paired-comparison preference listening test was conducted. This test compared the naturalness of synthesized speech generated from the models by MDL, SMAP, CVML, and CVSMAP over 100 sentences randomly selected from the test sentences. The model sizes for MDL, CVML, and CVSMAP were the automatically selected ones in these criteria. Because SMAP did not have the automatic stopping criterion, the model that gave the best test-set log likelihood was used. The listening test was carried out on Amazon Mechanical Turk [15]. To ensure that pairs of speech samples were played equally often in AB as in BA order, both orders were regarded as different pairs. Thus there were 2×100 evaluation pairs in the test. One subject could evaluate up to 40 pairs, they were randomly chosen and presented for each subject. Each pair was evaluated by three subjects. After listening to each pair of samples, the subjects were asked to choose their preferred one. Note that the subjects could select "No preference" if they had no preference.

Table 1 shows the preference test result. The differences were statistically significant at the 1% level by the paired t-test except the one between MDL and SMAP. It can be seen from the table that the use of CV significantly improved the naturalness of syn-

Table 1. Preference scores (%) between speech samples synthesized from the models clustered by the MDL, SMAP, CVML, and CVSMAP criteria.

MDL	SMAP	CVML	CVSMAP	No pref.
33.1	35.0	–	–	31.9
31.4	–	**39.2**	–	29.4
29.3	–	–	**40.1**	30.6
–	28.9	–	**39.5**	31.6
–	–	27.2	**31.7**	41.1

thesized speech in both with and without hierarchical regularization. Although the use of hierarchical priors slightly improved the naturalness, it was not as large as that by CV.

One problem with HMM-based speech synthesis is the quality of synthesized speech; the synthesized speech often sounds buzzy and muffled. One reason for this problem is oversmoothing caused by underfitting of the model to the data [16]. A possible way to reduce this problem is to increase the model size. As CVSMAP can select a larger model size, it is expected that it can reduce the oversmoothing problem. Figure 2 shows the running spectra generated from the models by MDL and CVSMAP with and without GV. It can be seen from the figure that the formant structure of spectra from CVSMAP without GV was slightly clearer than that from MDL without GV but the effect of GV was much larger than CVSMAP. It suggests that the oversmoothing problem mainly comes from the mismatch between the model and data rather than the size of the model.

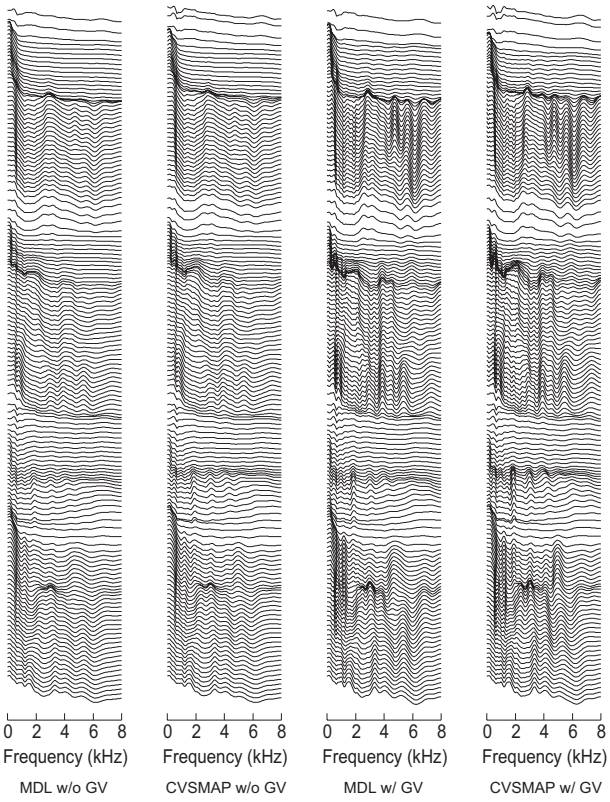

Fig. 2. Running spectra for a test sentence generated from the models clustered by the MDL and CVSMAP criteria with and without using GV. Note that the model by CVSMAP had about five times larger number of parameters than that by MDL.

4. CONCLUSIONS

This paper proposed an approach for decision tree-based context clustering based on cross validation and hierarchical priors. Combination of cross validation and hierarchical priors within decision tree-based context clustering offers better model selection and more robust parameter estimation than conventional approaches with no tuning parameters. Experimental results on HMM-based speech synthesis showed that the proposed approach achieved significant improvements in naturalness of synthesized speech over the conventional approaches. They also suggested that the oversmoothing problem in HMM-based speech synthesis comes from the model mismatch rather than the model size.

The proposed approach selects a larger model size than the conventional approaches. Future work includes compacting the model size based on the proposed criterion.

5. REFERENCES

[1] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002, CD-ROM Proceeding.

[2] J.J. Odell, *The use of context in large vocabulary speech recognition*, Ph.D. thesis, Cambridge University, 1995.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[4] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eurospeech*, 1997, pp. 99–102.

[5] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 365–381, 2006.

[6] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[7] I. Rogina, "Automatic architecture design by likelihood-based context clustering with cross validation," in *Proc. Eurospeech*, 1997, pp. 1223–1226.

[8] T. Shinozaki, "HMM state clustering based on efficient cross-validation," in *Proc. ICASSP*, 2006, pp. 1157–1160.

[9] Y. Zhang, Z.-J. Yan, and F.-K. Soong, "Cross-validation based decision tree clustering for HMM-based TTS," in *Proc. ICASSP*, 2010, pp. 4602–4605.

[10] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276–287, 2001.

[11] F. Flego and M.J.F. Gales, "Incremental predictive and adaptive noise compensation," in *Proc. ICASSP*, 2009, pp. 3837–3840.

[12] H. Zen, T. Toda, M. Nakamura, and T. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[13] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge Workshop*, 2008.

[14] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[15] "Amazon Mechanical Turk," `http://www.mturk.com/mturk/`.

[16] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.