Synthetic References for Template-based ASR using Posterior Features

Serena Soldo [†,‡], Mathew Magimai.-Doss [†], Hervé Bourlard [†,‡]

[†] Idiap Research Institute, Martigny, Switzerland
[‡]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{serena.soldo, mathew, bourlard}@idiap.ch

Abstract

Recently, the use of phoneme class-conditional probabilities as features (posterior features) for template-based ASR has been proposed. These features have been found to generalize well to unseen data and yield better systems than standard spectral-based features. In this paper, motivated by the high quality of current text-to-speech systems and the robustness of posterior features toward undesired variability, we investigate the use of synthetic speech to generate reference templates. The use of synthetic speech in template-based ASR not only allows to address the issue of in-domain data collection but also expansion of vocabulary. Using 75- and 600-word task-independent and speaker-independent setup on Phonebook database, we investigate different synthetic voices produced by the Festival HTS-based synthesizer trained on CMU ARCTIC databases. Our study shows that synthetic speech templates can yield performance comparable to the natural speech templates, especially with synthetic voices that have high intelligibility.

Index Terms: Speech recognition, template-based approach, posterior features, synthetic reference templates.

1. Introduction

In standard template-based Automatic Speech Recognition (ASR) approach [1], each test utterance is first transformed into a sequence of short-time spectral-based features and then compared against a set of reference templates, using Dynamic Time Warping algorithm, to find the best match. Recently, the use of phone class-conditional posterior probabilities estimated by an MultiLayer Perceptron (MLP) directly as speech features has been proposed [2, 3]. We refer to these features as *posterior features*. It was shown that, as a result of training of the estimator, posterior features are robust to undesired variability and can generalize well, thus yielding significantly better performance than standard spectral-based features using a fewer number of templates. Section 2 provides an overview of the template-based ASR framework using posterior features and summarizes our previous findings.

In this paper, we further investigate the robustness of posterior features to introduce the use of synthetic speech as reference templates. The high quality of the current Text-to-speech (TTS) systems, together with the property of the MLP to generalize to unseen speech/condition [3], suggests that it could be possible to build more flexible template-based ASR systems. The use of automatically generated speech could help overcome the disadvantages of collecting in-domain templates or the difficulties in expanding the dictionary. We illustrate the new framework and motivate this idea in Section 3.

We perform template-based isolated word recognition using posterior features on small vocabulary (75 and 600 words), task- and speaker-independent Phonebook corpus setup with one template per word. We generate templates using different voices trained with a HTS-based Text-To-Speech (TTS) system. Our investigations show that good quality synthetic speech can yield performance comparable to the use of natural speech templates. Section 4 presents the experimental setup and Section 5 the results. Section 6 provides an analysis of results. Finally, Section 7 concludes the paper discussing the work in a broader context.

This work was partly supported by the European Union under the Marie-Curie Training project SCALE, Speech Communication with Adaptive LEarning and the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2). The authors would like to thank their colleagues for their comments, inputs and suggestions.

2. Posterior template-based ASR

Formally, given a spectral-based feature vector, $\mathbf{x}$, and given a set of possible phoneme classes c_k with $k \in \{1, 2, ..., K\}$, the posterior features vector $\mathbf{y}$ is given by $\mathbf{y} = [P(c_1|\mathbf{x}), \ldots, P(c_K|\mathbf{x})]^{\mathrm{T}} = [y_1, \ldots, y_K]^{\mathrm{T}}$. As discrete distribution, the vector $\mathbf{y}$ has two properties: *a)* $y_k \in [0, 1], \forall k \in \{1, 2, ..., K\}$ and *b)* $\sum_{k=1}^{K} y_k = 1$.

In our previous work [3], different aspects of these features have been investigated in the context of template-base ASR, showing that these features generalize well to unseen data and yield better systems than standard spectral-based features.

The framework of a template-based ASR system using posterior features is shown in Figure 1.

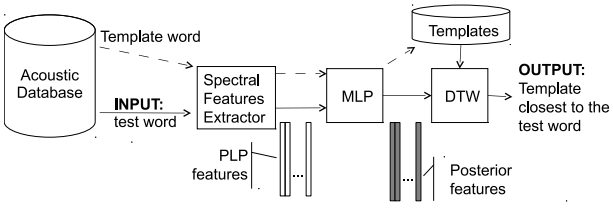

Figure 1: *Framework of a template-based ASR using posterior features*

The input speech signal is first transformed into a sequence of cepstral-based feature vectors. Each vector in the sequence (along with a temporal context) is then provided as input to an estimator and transformed into a posterior features vector. Different posterior features estimator were studied in [3] and it was found that, irrespective of the estimator, posterior features always yield better performance than spectral features. Specifically, MLP was found to yield consistently better systems.

In the training phase of such framework (dashed lines in figure 1), a number of reference templates are extracted from a database in the same domain as the test data. The templates are

transformed into a sequence of posterior features and stored in memory. In the test phase (continuous lines in figure 1), a test word is transformed into a sequence of posterior features and then compared to each template using the Dynamic Time Warping (DTW) algorithm. The decision making criterion provide as output the word corresponding to the best matching template.

In previous works, it was shown that the DTW algorithm can be redefined using a local distance measure (*local score*) that takes into account the probabilistic nature of posterior features, such as Bhattacharyya distance, Kullback-Leibler divergence, scalar product, cosine angle [2, 3]. These local distance measures were found to yield significantly better performance when compared to Euclidean distance. Furthermore, a local score based on Kullback-Leibler divergence, namely weighted symmetric Kullback-Leibler divergence (wSKL) was found to yield the best system. Briefly, if $\mathbf{y} = [y_1, \ldots, y_K]^{\mathrm{T}}$ denotes the posterior feature vector that belongs to the reference template and $\mathbf{z} = [z_1, \ldots, z_K]^{\mathrm{T}}$ denotes the posterior feature vector that belongs to the test template then wSKL is computed as:

$$wSKL(\mathbf{y}, \mathbf{z}) = w_{\mathbf{y}} \cdot KL(\mathbf{y}, \mathbf{z}) + w_{\mathbf{z}} \cdot RKL(\mathbf{y}, \mathbf{z}) \quad (1)$$

where,

$$KL(\mathbf{y}, \mathbf{z}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{z_k},$$

$$RKL(\mathbf{y}, \mathbf{z}) = \sum_{k=1}^{K} z_k \log \frac{z_k}{y_k},$$

$$w_{\mathbf{y}} = \frac{\frac{1}{H(\mathbf{y})}}{(\frac{1}{H(\mathbf{y})} + \frac{1}{H(\mathbf{z})})}, \quad w_{\mathbf{z}} = \frac{\frac{1}{H(\mathbf{z})}}{(\frac{1}{H(\mathbf{y})} + \frac{1}{H(\mathbf{z})})}$$

$H(\mathbf{y})$ is the entropy of $\mathbf{y}$, and $H(\mathbf{z})$ is the entropy of $\mathbf{z}$.

3. Synthetic References, Posterior Features and Template-based ASR

In the framework described in the previous section, the templates were still extracted from data of the same domain as the test data. Here, we study the possibility of using templates which are domain-independent and are generated using a TTS system. This would eliminate the issues related to in-domain data collection or vocabulary expansion. This new framework is illustrated in Figure 2.

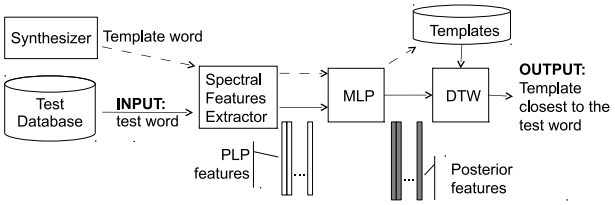

Figure 2: *Framework of the template-based ASR using posterior features and synthetic templates*

The use of synthetic templates has been already suggested in the past. In particular, in [4] the authors proposed a recognition system in which a speech production system (rule-based TTS system) is used to generate a number of synthetic reference templates that are matched to the input test utterance at spectral level. Besides overcoming the problem with data collection, the authors cite other reasons for using this approach. In particular, they highlight that synthetic speech can be modified to match the voice of the current speaker and it can also be used to exploit knowledge that is available about natural speech such as duration and the context of a word. Moreover, the perfect consistency of synthetic speech may be used to improve separation between words such as *stalagmite* and *stalactite* that have phonetically identical parts. However, the results of their experiment were far from competitive with systems based on natural speech templates. The failure of that work was mainly ascribed to the low quality of the voices produced by the rule-based TTS system and, thus, the lack of similarity between synthetic and natural speech.

Recently, new models for TTS has been proposed [5] and the quality of the synthetic voices has considerably increased. In these new approaches, the representation of the speech signal usually also includes information about the spectral envelope of the speech signal [6]. On the other hand, it has been shown that the MLP tends to learn information about the spectral envelope of the speech signal [7]. This suggests that MLPs could estimate reliable posterior features for synthetic speech as well. As first step, here we intend to investigate this framework in its simplest form, without introducing any kind of adaptation. In the following section, we present the experimental setup and details about the different components of the system.

4. Experimental Setup

We use the Phonebook speech corpus for speaker-independent task-independent isolated word recognition. This corpus contains US English read telephone speech. The test set consists of 8 subsets of utterances, each containing 75 words uttered on average by 11 or 12 speakers once. For more details about the composition of this dataset, the reader may refer to [8].

We perform our experiments on two different tasks:

- 75-word task: the recognition is performed on each of the 8 subset (75-word lexicon each) separately and the average word error rate is presented as result.
- 600-word task: the 8 test subsets are merged to setup a task with 600 words lexicon.

In this work, we use exactly same framework as in [2][3], where one random utterance of each word was extracted from the test set and used as natural speech reference template. There are two voices, namely, one female (denoted as natural voice 1) and one male (denoted as natural voice 2).

Though this work focuses on template-based ASR using posterior features, for sake of completeness we also report studies with standard spectral-based feature. More precisely, using 39-dimensional PLP cepstral feature vector ($c_0 - c_{12} + \Delta + \Delta\Delta$).

Text-To-Speech system

The synthetic reference templates were generated using Festival Speech Synthesis System [9]. We used *off-the-shelf* HMM-based Speech Synthesis System (HTS) voices, trained using the CMU ARCTIC databases [10]. These databases consist of phonetically balanced sentences selected from out-of-copyright texts recorded using a microphone in a sound proof room. Among the different voices available, we used two US English male voices (*BDL* and *RMS*) and two US English female voices (*SLT* and *CLB*). For more details about the training system the reader may refer to [11]. In our experiments, each of the four synthetic voices has been used to produce one utterance of each word in the dictionary.

Posterior features estimation

We estimate posterior features using the MLP that yielded the best system in our previous work [3]. This MLP was trained with 232 hours of conversational telephone speech. The input to the MLP is a vector of 39-dimensional PLP features along with a temporal context of 90ms. The MLP has

5000 hidden units and 45 output units, each corresponding to a context-independent phoneme.

In the case of synthetic speech, the speech was down sampled from 16 kHz to 8 kHz and posterior features were extracted without performing any kind of adaptation on the MLP.

Local Scores

In case of PLP features, Euclidean distance is used as local score in the DTW algorithm to compare two frames of feature vectors.

For posterior features, we use weighted symmetric KL-divergence defined earlier in Equation (1) which was found to yield the best system in previous studies [2, 3].

We use the same local scores for both natural and synthetic speech.

5. Results

Figures 3 and 4 present the results obtained for both 75- and 600-words task using PLP features and posterior features respectively. The performances are expressed in terms of word accuracy. *Natural* denotes the system with natural voices and *Synthetic* denotes the system with synthetic voices.

Using PLP features, on 75-words task, the best result with natural speech templates is 65.1% word accuracy, whereas the best results with synthetic speech templates is 67.8%. On 600-word task, the best result with natural speech templates is 41.0% word accuracy, whereas the best results with synthetic speech templates is 48.5%. Using posterior features, on 75-word task, the best result obtained with natural speech templates is 98.8% word accuracy, whereas the best results with synthetic speech templates is 98.2%. On 600-words task, the best result obtained with natural speech templates is 94.8% word accuracy, whereas the best results obtained with synthetic speech templates is 94.7%. In case of synthetic voices, *RMS* voice yields the best performance for both PLP and posterior features.

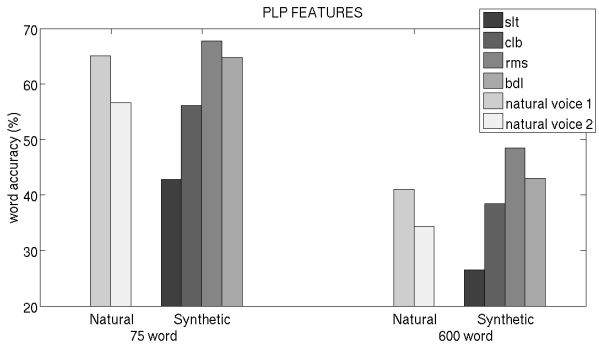

Figure 3: *Word accuracy for PLP features on 75- and 600-words tasks using different voices.*

Overall, it can be observed that the use of synthetic speech templates can provide results comparable to the natural speech templates. Interestingly, PLP features also show such trend, supporting the suggestion put forward in [4] on the feasibility of this idea provided good quality synthetic speech. However, posterior features appear to be more robust to speaker variations than PLP features. Finally, it can be observed that not all the synthetic voices perform in the same way. We elucidate this aspect in the next section.

6. Analysis

To gain a better insight into the behavior of our system in case of synthetic templates, we tried to relate its performance to the quality of the synthetic speech. An evaluation of the four synthetic voices in terms of naturalness and intelligibility is provided in [13]. Figure 5 shows the relation between the system word accuracy using posterior features and the naturalness and the intelligibility, respectively. Figure 6 shows a similar comparison using PLP features.

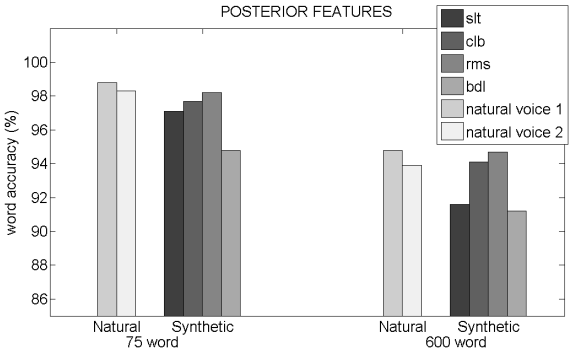

Figure 4: *Word accuracy for posterior features on 75- and 600-words tasks using different voices. The hybrid HMM/MLP system on this task yields 98.8% and 96.0% word accuracy, respectively [12].*

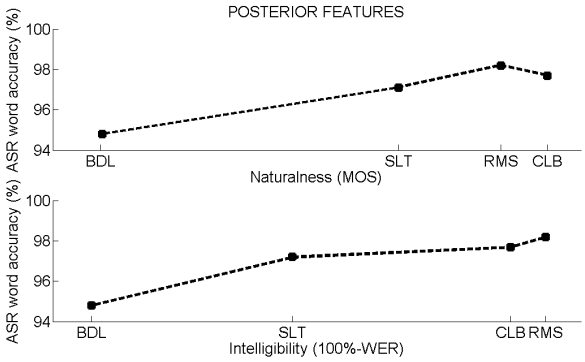

Figure 5: *Comparison of subjective evaluations of the voices and word accuracy of the system using posterior features. The average MOS values for the four voices are:* RMS *3.00* - CLB *3.03* - SLT *2.94* - BDL *2.75. The average WER values are:* RMS *14.0%* - CLB *15.0%* - SLT *21.2%* - CLB *26.9%*

It can be observed that, when posterior features are used, there is a clear positive relation between the performance of the system and the quality of the synthetic speech, especially in case of intelligibility. In the case of PLP features, it appears that there is no clear relation between the system accuracy and either of the two subjective measures.

As part of the analysis, we also investigate a scenario where several templates per word are available. Once fixed the number of templates per word (between 1 and 4), we can build different scenarios using all possible combinations of the four synthetic voices. Figure 7 shows the results obtained on 75-word task varying the number of templates. To provide a complete picture of the results, for each number of templates we show the average performance over all the scenarios together with the best and worst performance. The dashed line indicates the best result achieved using natural speech templates (corresponding to the 2-template scenario in [3]). It can be observed that the average behavior for each scenario tends to converge to the best result obtained using 1-template scenario (corresponding to the use of *RMS* voice). This suggests that there is no complementarity be-

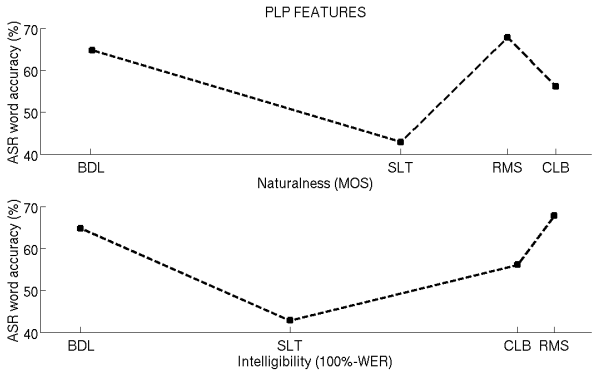

Figure 6: *Comparison of subjective evaluations of the voices and word accuracy of the system using PLP features*

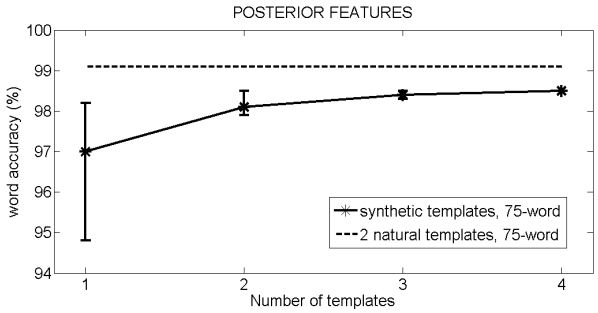

Figure 7: *Word accuracy of the system increasing the number of synthetic speech templates. The dashed line corresponds to the best results obtained using the natural speech templates (2-template scenario).*

tween voices. From an analysis of the output of the recognition system, it emerges that only combining the templates produced by the *RMS* voice with those produced by the *CLB* voice corresponds to a small improvement of the performance, whereas the combination with the other voices provides the same performance as when only *RMS* voice is used. In other words, the templates corresponding to the *RMS* voice are, in most of the cases, the best match. However, combining templates produced by two voices with lower quality, in general, provides an improvement in the performance. This aspect needs further investigation and is part of our future work.

7. Summary and Conclusion

In this work we investigated the use of synthetic references for template-based automatic speech recognition using posterior features. In our studies, it was found that without performing any kind of adaptation on the domain/task-independent posterior feature estimator (i.e., MLP) the system yields comparable performance to the use of natural speech templates. This can be attributed to the ability of current TTS systems to generate high quality synthetic speech, the ability of MLP to robustly estimate posterior features, and the use of appropriate local score.

This finding is not only relevant to template-based ASR, but also to HMM-based ASR. Indeed, as shown already shown in [14, 2], the use of posterior features, in conjunction with local scores is yielding a common theoretical framework. In the latter case, the system is referred to as Kullback-Leibler divergence based HMM (KL-HMM). Thus, it may be possible to use synthetic speech for training KL-HMMs.

In conclusion, our work shows that synthetic speech together with posterior features can be exploited to develop flexible template-based ASR system.

Our future work will scrutinize several issues, including:

- Higher quality TTS: our analysis indicates that just addition of multiple voices may not improve the performance of the system. One approach would be to look for high quality synthetic speech.
- Better posteriors: the posterior feature estimator could be adapted using hierarchical MLP-based approach, where the second MLP trained on top of the first MLP has the ability to learn confusions present at the output of the first MLP and also learn phonotactic constraints [12], to improve the performance on low quality synthetic voices.
- Natural speech templates: investigating (and exploiting) the complementarity between natural speech and synthetic speech templates.
- More natural HTS voices: in the work reported here, HTS voices were trained on CMU ARCTIC database which was collected using a microphone in a sound proof room. We intend to investigate if the findings are generalizable to voices trained on other corpus, such as Wall Street Journal.

8. References

[1] L. Rabiner and H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[2] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Posterior Features Applied to Speech Recognition Tasks with User-Defined Vocabulary," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[3] S. Soldo, M. Magimai.-Doss, J. Pinto, and H. Bourlard, "Posterior Features for Template-based ASR," in *Proc. of ICASSP*, Prague, Czech Republic, 2011.

[4] M. Blomberg, R. Carlson, K. O. E. Elenius, B. Granstrm, and S. Hunnicutt, "Word recognition using synthesized templates," Department of Speech Communication and Music Acoustics, KTH, Sweden, Tech. Rep. STL-QPSR 2-3/1988, 1988.

[5] S. King, "An introduction to statistical parametric speech synthesis," *Sādhanā*, vol. 36, no. 5, pp. 837–852, 2011.

[6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[7] J. Pinto, G. S. V. S. Sivaram, H. Hermansky, and M. Magimai.-Doss, "Volterra series for analyzing mlp based phoneme posterior probability estimator," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[8] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on 'PhoneBook' and Related Improvements," in *Proc. of ICASSP*, Munich, Germany, 1997.

[9] A. Black and K. Lenzo, "Building voices in the festival speech synthesis system," http://festvox.org/bsv, 2000.

[10] J. Kominek and A. W. Black, "CMU Arctic Databases for Seech Synthesis," Language Technologies Institute, Carnegie Mellon University, Tech. Rep. CMU-LTI-03-177, 2003.

[11] 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. of Blizzard Challenge 2008*, 2008.

[12] J. Pinto, M. Magimai.-Doss, and H. Bourlard, "MLP Based Hierarchical System for Task Adaptation in ASR," in *Proc. of ASRU*, Merano, Italy, 2009.

[13] C. L. Bennett, "Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005," in *Proc. of Interspeech*, Lisbon, Portugal, 2005, pp. 105–108.

[14] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Using KL-based Acoustic Models in a Large Vocabulary Recognition Task," in *Proc. of Interspeech*, Brisbane, Australia, 2008.