ADAPTATION OF A SPEECH RECOGNIZER FOR SINGING VOICE

Annamaria Mesaros, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 1, 33720, Tampere, Finland
email: annamaria.mesaros@tut.fi, tuomas.virtanen@tut.fi

ABSTRACT

This paper studies the speaker adaptation techniques that can be applied for adapting a speech recognizer to singing voice. Maximum likelihood linear regression (MLLR) techniques are studied, with specific details in choosing the number and types of transforms. The recognition performance of the different methods is measured in terms of phoneme recognition rate and singing-to-lyrics alignment errors of the adapted recognizers. Different methods improve the correct recognition rate with up to 10 percentage units, compared to the non-adapted system. In singing-to-lyrics alignment we obtain a best of 0.94 seconds mean absolute alignment error, compared to 1.26 seconds for the non-adapted system. Global adaptation was found to provide the most improvement in the performance, but small further improvement was obtained with regression tree adaptation.

1. INTRODUCTION

Recognition of lyrics has a large potential in music information retrieval (MIR). Until now, information retrieval from singing has focused on melodic information, ignoring the lyrics. The lyrics are an important aspect of music since they carry the semantic information. Early attempts to perform lyrics recognition using a large-vocabulary speech recognizer were somewhat successful on pure singing voice [1, 2], but the performance is still limited.

A lyrics recognition system can be developed based on principles used in automatic speech recognition. Building a phonetic recognizer requires a large database of examples, multiple annotated recordings of phonetically balanced sentences. There is no such database available for singing, therefore for singing recognition we have to resort to models trained on speech data. In automatic speech recognition it is possible to build improved acoustic models by adapting a speaker-independent model to a specific speaker. Using a small amount of data from the target speaker, the speaker-independent set of models can be adapted to better fit the characteristics of the target speaker. The same approach can be used in building a recognition system for the singing voice [3].

Our previous work tackles the automatic alignment of polyphonic music with textual lyrics in English, by using a speech recognizer [4] . Using standard maximum-likelihood linear regression (MLLR) speaker adaptation technique, the monophone models trained on speech data were adapted to clean singing voice characteristics [3, 4]. The adaptation was found to be beneficial in the alignment.

This work represents a more thorough study of the speaker adaptation technique, looking into different possibilities of generating the linear transforms used in the adaptation, in order to obtain better models for singing phonemes. The paper is organized as follows: Section 2 presents the principles and basic alternatives of supervised adaptation. Section 3 presents the specific details we chose for adapting the speech-trained models to singing and the singing-to-lyrics alignment application used for testing the adapted models. Sections 4 and 5 present simulation experiments, discussion and conclusions.

2. HMM PHONETIC RECOGNIZER

Singing recognition can be done using a phonetic hidden Markov model (HMM) recognizer, which is the standard technique in automatic speech recognition. In HMM based speech recognition it is assumed that the observed sequence of speech vectors is generated by a hidden Markov model. An HMM consists of a number of states with associated observation probability distributions and a transition matrix defining transition probabilities between the states. Phoneme-level recognizers use typically a 3 state left-to-right HMM to represent each phoneme. The emission probability of each state is modeled by a Gaussian mixture model (GMM). In the training process, the transition matrix and the means and variances of the Gaussian components in each state are estimated to maximize the likelihood of the observation vectors in the training data.

2.1 MLLR adaptation

Each speaker has slightly different characteristics in speaking, but construction of a speaker-dependent speech recognizer is not always feasible. It is possible to improve the performance of a speaker-independent recognizer on a target speaker by using speaker adaptation. Speaker adaptation methods use a smaller amount of the target speaker voice material to adapt the recognizer parameters in order to reduce the mismatch between the trained models and the speaker data. The speaker-specific data is referred to as adaptation data. Maximum likelihood linear regression (MLLR) [5] computes a set of transformations for the means and variances of a Gaussian mixture HMM system. The transformations shift the component means and variances of the initial system so that the resulting HMM is more likely to generate the adaptation data.

The transformation matrix used to give a new estimate of the adapted models is obtained by solving a maximization problem using the Expectation-Maximization technique. The adaptation of the mean vector of each Gaussian component is carried out using a transformation matrix $\mathbf{W}$ as:

$$\hat{\boldsymbol{\mu}} = \mathbf{W}\boldsymbol{\xi} \tag{1}$$

where $\mathbf{W}$ is the $n \times (n+1)$ transformation matrix (n is the dimensionality of the data) and $\boldsymbol{\xi}$ is the extended mean vec-

tor $\boldsymbol{\xi} = [1 \quad \boldsymbol{\mu}]^T$, with $\boldsymbol{\mu}$ being the mean vector of a mixture component. $\mathbf{W}$ can be decomposed into $\mathbf{W} = [\mathbf{b} \quad \mathbf{A}]$, where $\mathbf{A}$ represents the transformation matrix and $\mathbf{b}$ is a bias.

In the MLLR adaptation of the mean (MLLRMEAN), the adapted models will be characterized by mean $\hat{\boldsymbol{\mu}}$ in Equation 1. In MLLRMEAN, the covariance matrices of the models are not transformed.

The constrained MLLR (CMLLR) applies the same transformation matrix to the means and the covariances of the models, so that the adapted covariance matrix $\hat{\boldsymbol{\Sigma}}$ of a mixture component is obtained as:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T, \tag{2}$$

where $\boldsymbol{\Sigma}$ is the original covariance matrix of the mixture component.

2.2 Number of adaptation transforms

Iterative adaptation steps and multiple transforms can be used to modify the means and variances of the Gaussian components of each state in the HMM. Several components can be grouped into relevant categories called *base classes*. The classes can be defined based on the acoustic similarity of the models, states or Gaussian components. The components associated to one base class will share the same transform.

Depending on the amount of available data it is important to determine the appropriate set of transforms. If a small amount of adaptation data is available, the simplest form is to generate a global transform [6]. A global transform is applied to every Gaussian component in the model set, no matter the phoneme model or the state. After this, a second iteration can be done, with increased number of transforms . Each transform will be more specific and applied to certain Gaussian components, grouped in base classes.

A dynamic way of specifying the base classes uses a regression tree to group the Gaussians in the initial model set. The regression tree is constructed such that it clusters together mixture components that are close in the acoustic space [6]. The regression tree is built using the original model set, and thus it is independent of the adaptation data. In this definition, the leaves of the tree are the base classes and they specify the component groupings. Each Gaussian component of each state of each phoneme model belongs to one particular base class, making this a very detailed grouping of the model parameters.

3. SINGING RECOGNITION

Speech and singing voice sounds have many properties in common because they convey the same kind of semantic information and originate from the same production physiology. In singing, however, the intelligibility is often secondary to the intonation and musical qualities of the voice. Vowels are sustained much longer in singing than in speech and independent control of pitch and loudness over a large range is required. The dynamic range is greater in singing than in speech, and also the fundamental frequency variations of singing are of about 2 octaves for an average trained singer.

In speech recognition, feature-space transformations are used to compensate for the difference in pitch. The method called vocal tract normalization (VTN) was found to be equivalent to a constrained MLLR transformation in the model space, with the transformation matrix being controlled by VTN frequency warping parameter [7]. Singing at high pitch includes fundamental frequencies greater than the first formant, determining a trained singer to adjust the formants position, with a small loss of perceived vowel quality [8]. Depending on the singer's skills, the vowels in singing may have different spectral characteristics than vowels in speech, and this cannot be compensated by VTN.

Our scope is to investigate adaptation of a speech recognizer to singing voice using different grouping of phonemes into classes and test the recognition performance of the resulting adapted models. For this, we trained a HMM speech recognizer consisting of 39 monophone models plus silence and short pause models.

As features, we used 13 mel-frequency cepstral coefficients plus delta and acceleration coefficients, calculated in 25 ms frames with a 10 ms hop between adjacent frames. A left-to-right HMM with 3 states is used to represent each phoneme. The silence model is a fully connected HMM with 3 states and the short pause is a one-state HMM tied to the middle state of the silence model. The system was implemented using HTK [1].

The training was done using the CMU ARCTIC speech database [2]. For adaptation we used a singing voice database, consisting of 49 fragments of 12 pop songs, ranging from 20 to 30 seconds. There are 19 male and 30 female voice fragments. The phonetic transcription of the fragments was available, so that adaptation could be done in a supervised manner.

3.1 Singing adaptation

In speech analysis, the phonemes are usually classified into 7 broad categories: monophtongs, diphtongs, approximants, nasals, fricatives, plosives, affricates. In singing, the rate of voiced sounds can be up to 95%, compared to about 60% in speech [8], because the voiced sounds carry the musical information. Considering this, we can think about singing as composed from two types of sounds: vowels and consonants. A third way of classifying the singing sounds is considering each vowel as a separate base class, and grouping the rest of phonemes into consonants into approximants, nasals, fricatives, plosives and affricates. One more base class can be added to model the silence and short pause nonspeech events.

We devise two adaptation strategies consisting of two iterations each. The first strategy is by using a global transform in the first iteration. The second iteration uses base classes obtained by means of a regression tree, which is data-driven. The base classes defined at this stage are based on the acoustic similarity of mixture components. We test MLLREAN and CMLLR adaptation to determine which type of adaptation and what number of base classes (3, 8 or 22) is better suited for the adaptation. Different adaptation methods are denoted as GT*i*M or GT*i*C, where G is the global transform, T*i* is a tree-based transform with *i* base classes and M or C denote MLLRMEAN and CMLLR, respectively.

In the second adaptation strategy, we define base classes at phoneme-level, according to the phoneme classifications presented above, to replace the global transform for the first iteration. We obtain 3, 8 or 22 transforms in the place of a

[1] The Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/HTK

[2] CMU ARCTIC databases for speech synthesis, http://festvox.org/cmu_arctic/

Table 1: *Different adaptation methods and the number of classes in the two adaptation iterations. The first iteration: global or phoneme-dependent base classes, the second iteration: mixture-level base classes, based on the acoustic similarities*

method	first iteration	second iteration	adaptation type
GT3M	1	3	MLLRMEAN
GT3C	1	3	CMLLR
GT8M	1	8	MLLRMEAN
GT8C	1	8	CMLLR
GT22M	1	22	MLLRMEAN
GT22C	1	22	CMLLR
G3	3	not used	CMLLR
G8	8	not used	CMLLR
G22	22	not used	CMLLR
G3T8	3	8	CMLLR
G8T8	8	8	CMLLR
G22T8	22	8	CMLLR

single (global) transform, resulting in the adaptation methods G3, G8 and G22 (see Table 1). The second iteration uses 8 base classes obtained by means of the regression tree, using the component-wise acoustic similarities. The systems are denoted as GjT8, having j phoneme-level base classes and tree-based transform with 8 base classes. The combinations of user defined and data-driven number of classes and corresponding adapted systems is summarized in Table 1.

3.2 Audio-to-lyrics alignment

In addition to phooneme recognition, the previously developed application, audio to lyrics alignment [4] can be used to test the adapted models. We align the singing voice from a polyphonic audio with the corresponding textual lyrics. The audio file is preprocessed to separate the vocal line from the polyphonic mixture.

For separating the singing from the polyphonic signal, we use the system proposed in [9]. The system first estimates the time-varying pitch of the singing. By assuming perfect harmonicity, it generates a binary mask which indicates the presence of singing in each time-frequency point of the spectrogram. The method learns a model for the accompaniment using non-negative matrix factorization in the non-vocal time-frequency regions, and subtracts the accompaniment model from the singing regions. The time-domain signal corresponding to the singing is generated by using phases of the mixture signal, inverse discrete Fourier transform, and overlap-add. In [9] the system was found to produce better separation quality than reference system based on sinusoidal modeling or basic binary masking.

After separating the vocal line, we extract features of the singing voice. An additional noise model trained on instrumental fragments is used to account for occasional distorted instrumental fragments that remain in the separated signal.

The lyrics text is preprocessed to obtain a sequence of words with optional silence, pause and noise between them. The transcription from words to phonemes is done using the CMU pronouncing dictionary. The features extracted from the separated vocals are aligned with the obtained string of phonemes, using Viterbi forced alignment. The alignment can be viewed as a special case of recognition, with the possible paths in the Viterbi search algorithm restricted to the one given by the phoneme sequence from the text.

4. SIMULATION EXPERIMENTS

We evaluate the performance of the adaptation methods by the phoneme recognition rate of singing and the accuracy of singing-to-lyrics alignment. The phoneme recognition task uses no language model, which means that any phoneme can follow any other phoneme. Correctness and accuracy of the recognition are defined in terms of the number of substitution errors S, deletion errors D and insertion errors I, reported to the total number of tested instances, N:

$$correct[\%] = \frac{N-D-S}{N} \times 100$$

$$accuracy[\%] = \frac{N-D-S-I}{N} \times 100$$

The alignment performance is measured using the mean absolute error in alignment at the start and end of each line in the lyrics, on a number of 100 sections from 17 songs, manually annotated for reference.

For testing the phoneme recognition rate of the adapted systems we used a 5-fold experimental setup on the acoustic material described in Section 3. The test data was approximately one fifth of the entire singing material, while the rest was used for adaptation. For testing the alignment performance, we adapted the systems with the entire singing material, consisting of a total of 4770 phonemes.

The singing phoneme recognition rate of the non-adapted recognizer is 33.29% with an accuracy of -6.4%. The accuracy of the recognition is very low due to a large number of insertion errors. The mean absolute error for the alignment task is 1.26 seconds.

4.1 MLLRMEAN and CMLLR

Our first goal was to test which of the two adaptation methods yields a better result: adaptation of the means only or the constrained adaptation (means and variances transformed with the same matrix). For this, the adaptation was performed using one global transform and different number of classes tree-based adaptation. The regression trees were generated on the speech material. In this task we included all the singing data in the adaptation and report the phoneme recognition rate on the adaptation data.

Table 2 lists the phoneme recognition results of MLLRMEAN and CMLLR adapted systems with one global transform and a tree-base transform with 8 base classes, in the 5-fold setting, compared to the performance of the non-adapted system. Both methods improve the phoneme recognition rate and accuracy in comparison with the non-adapted models, CMLLR being slightly better.

The systems adapted with different numbers of base classes in the second iteration were tested in the alignment task. The results are presented in Table 3. While the differences between the differently adapted systems in the the phoneme recognition were small, we see quite a large variation in the alignment results. The minimum alignment error is obtained by the models adapted using CMLLR with one global class and tree with 8 base classes. Using 22 classes

Table 2: *Phoneme recognition results of GT8M and GT8C adapted systems on the test set and adaptation set*

method	test set correct / acc [%]	adaptation set correct / acc [%]
non-adapted	33.29 / -6.40	-
GT8M	40.86 / 20.85	62.81 / 47.52
GT8C	41.31 / 18.94	62.90 / 47.41

Table 3: *Average singing to lyrics alignment errors for GTiM and GTiC adapted systems*

method	alignment error
non-adapted	1.26 s
GT3M	1.24 s
GT3C	1.14 s
GT8M	1.12 s
GT8C	1.02 s
GT22M	1.34 s
GT22C	1.16 s

in the regression tree adaptation reduces the alignment performance of the system. For 22 base classes there is not enough adaptation data for estimating a reliable transformation of each base class. In the alignment, CMLLR performed substantially better than MLLR.

4.2 Phoneme-level base classes

In the second adaptation experiment we replace the global transform with classes defined according to phoneme classifications, constructing phoneme-level base classes. We defined 3, 8 and 22 base classes for the first iteration in the adaptation. Based on the previous experiment, we decided to use the CMLLR adaptation only. In the second iteration, we used a tree based transform with 8 base classes.

The results of the phoneme recognition task are presented in Table 4. The systems G3 and G8 are very close to eachother in the phoneme recognition rate, both on the testing and on the adaptation data. After the second iteration, they continue to stay at similar performance level, as G3T8 and G8T8, with G3T8 having better accuracy. The performance of G22 and G22T8 is lower than of the other systems because there is not enough adaptation data for estimating a reliable transformation of each of the 22 base classes. The first adaptation iteration accounts for the largest improvement, while the second iteration improves the performance of all the methods by about one or two percentage units.

The mean absolute error of the alignment performance is presented in Table 5. The G3T8 and G8T8 systems achieve under 1 second mean absolute error, the minimum of 0.94 seconds belonging to G8T8. When all the singing data was used for adaptation, the recognition rate of the G*j*T8 adapted systems on the adaptation data is around 73% correct, with over 47% accuracy. This is a 10% increase compared to the 62% average of the 5-fold experiments. Thus, the amount of adaptation data is important.

Table 4: *Phoneme recognition results of Gj and GjT8 adapted systems on the test set and adaptation set*

method	test set correct / acc [%]	train set correct / acc [%]
G3	40.39 / 19.90	60.59 / 46.00
G8	40.29 / 18.72	60.74 / 45.12
G22	38.39 / 18.67	57.47 / 40.97
G3T8	41.35 / 20.01	62.56 / 47.26
G8T8	41.05 / 18.95	62.45 / 46.98
G22T8	40.68 / 19.53	62.02 / 46.44

Table 5: *Average singing to lyrics alignment errors for Gj and GjT8 adapted systems*

method	alignment error
G3	1.27 s
G8	1.31 s
G22	1.31 s
G3T8	0.97 s
G8T8	0.94 s
G22T8	1.07 s

5. CONCLUSIONS

This paper investigated MLLR adaptation of a speech recognizer to singing voice. We tested the mean and the constrained adaptation with different number of classes and transforms at phoneme level or mixture level. CMLLR provided slightly better phoneme recognition rates and clearly better alignment accuracy.

A single global transform was previously found to improve the performance, but changing it into a more specific one by taking into account phoneme properties also improved the results. Two adaptation iterations are needed because the first one considers phoneme-level transforms, whereas the second one captures details about the similarity of the initial model mixture components.

REFERENCES

[1] A. Loscos, P. Cano, and J. Bonada, "Low-delay singing voice alignment to text," in *Proceedings of the International Computer Music Conference (ICMC)*, 1999.

[2] T. Hosoya, M. Suzuki, A. Ito, and S. Makino, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," in *Proceedings of the 6th International Conference on Music Information Retrieval ISMIR*, 2005.

[3] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in *ISM '06: Proceedings of the Eighth IEEE International Symposium on Multimedia*, Washington, DC, USA, 2006.

[4] A. Mesaros and T. Virtanen, "Automatic alignment of music audio and lyrics," in *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, 2008.

[5] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, 1996.

[6] P. Woodland, M. Gales, D. Pye, and S. Young, "Broadcast news transcription using HTK," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Washington DC, USA, 1997.

[7] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, 2005.

[8] J. Sundberg, *The Science of Singing Voice*, Northern Illinois University Press, 1987.

[9] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition SAPA*, Brisbane, Australia, 2008.