Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication xxx (2013) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Eigentrigraphemes for under-resourced languages

Tom Ko, Brian Mak *

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Abstract

Grapheme-based modeling has an advantage over phone-based modeling in automatic speech recognition for under-resourced languages when a good dictionary is not available. Recently we proposed a new method for parameter estimation of context-dependent hidden Markov model (HMM) called *eigentriphone modeling*. Eigentriphone modeling outperforms conventional tied-state HMM by eliminating the quantization errors among the tied states. The eigentriphone modeling framework is very flexible and can be applied to any group of modeling unit provided that they may be represented by vectors of the same dimension. In this paper, we would like to port the eigentriphone modeling method from a phone-based system to a grapheme-based system; the new method will be called *eigentrigrapheme modeling*. Experiments on four official South African under-resourced languages (Afrikaans, South African English, Sesotho, siSwati) show that the new eigentrigrapheme modeling method reduces the word error rates of conventional tied-state trigrapheme modeling by an average of 4.08% relative.

Keywords: Eigentriphone; Eigentrigrapheme; Under-resourced language; Grapheme; Regularization; Weighted PCA

1. Introduction

In the past, research efforts on automatic speech recognition (ASR) are highly focused on the most popular languages such as English, Mandarin, Japanese, French, German, ... etc. in the developed countries. The remaining languages in the world, lacking audio and language resources, are considered under-resourced languages. Usually the phonetics and linguistics of these languages are not well studied either, thus the development of human language technologies for these languages is greatly hindered. Nonetheless, some of these under-resourced languages are spoken by a large population. For example, Vietnamese is spoken by about 80 million people, and Thai is spoken by 60 million people. It is not difficult to see that real-life ASR applications for these languages have a great potential. One major obstacle in developing an ASR system for under-resourced languages is the availability of data. It is usually costly and labor-intensive to create audio recordings and their human annotated transcriptions, and make linguistic analyses for languages. As a consequence, it is both academically intriguing and commercially attractive to look for economically more efficient and faster ways to create human language technologies for the under-resourced languages.

In order to reduce the amount of annotated audio data for training the acoustic models of a new target language, cross-lingual (Ogbureke et al., 2010; Le and Besacier, 2009) and multi-lingual (Kohler et al., 1996) acoustic modeling techniques have been developed. The rationale behind these techniques is that an acoustic model may be ported to or adapted from some other high-resourced languages, and only a relatively small amount of training data is required for the target language. A key step for these cross-lingual or multi-lingual techniques to work is to figure out a good mapping between phonemes across the languages. This can be done either using a knowledge-based (Beyerlein et al., 1999) or data-driven approach (Kohler et al., 1996). In the data-driven approach, the similarities between sounds can be measured by various distance measures such as

* Corresponding author. Tel.: +852 2358 7012.
E-mail addresses: tomko@cse.ust.hk (T. Ko), mak@cse.ust.hk (B. Mak).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.01.010

confusion matrix (Beyerlein et al., 1999), entropy-based distance (Kohler et al., 1996) or Euclidean distance (Sooful et al., 2001). The approach is further improved when the underlying model is more compactly represented. A notable example is the use of subspace Gaussian mixture model (SGMM) (Povey et al., 2010) in multi-lingual ASR (Burget et al., 2010; Lu et al., 2011). Another research direction is heading to making linguistic analysis of a target language easier and faster. Deducing the phone set and preparing the pronunciation dictionary for a new language usually require native linguistic experts. This process is expensive and time-consuming, and is even more so for non-native developers. One way to partially automate the development of a pronunciation dictionary is to first prepare a small primary dictionary manually, and then use it to bootstrap a large dictionary by applying grapheme-to-phoneme conversion (Meng et al., 1996; Bellegarda et al., 2003; Davel et al., 2004; Andersen et al., 1996). However, the performance of the final dictionary highly depends on the quality of the primary one. If the primary dictionary is not rich enough and does not cover all the implicit grapheme-to-phoneme relations in the language, the performance of the overall system will be hampered.

On the other hand, there is a simple solution to the creation of the phone set and pronunciation dictionary for an under-resourced language: there is no need to develop them if graphemes instead of phonemes are adopted as the acoustic modeling units. In grapheme modeling (Schukat-Talamazzini et al., 1993; Kanthak et al., 2002; Charoenpornsawat et al., 2006; Le and Besacier, 2009), each word in the "pronunciation dictionary" is simply represented by its graphemic transcription according to its lexical form. According to Daniels and Bright (1996), there are six types of writing systems in the world: logosyllabary, syllabary, abjad, alphabet, abugida, and featural. Many languages that use the alphabet writing system are suitable for grapheme acoustic modeling, and their grapheme set is usually selected to be the same as their alphabet set (Schukat-Talamazzini et al., 1993).

The performance of grapheme modeling in ASR is sensitive to the languages. For example, it works better than phone modeling in Spanish but worse than phone modeling in English and Thai (Stuker, 2009). The reason is that the pronunciation of English has developed away from its written form over time, whereas Thai has some complex rules that map its writing to the pronunciation. There are techniques that improve grapheme modeling; for example, in Charoenpornsawat et al. (2006), a text normalization scheme was applied on Thai graphemes to improve the performance of a Thai ASR system. There are also works on multi-lingual grapheme modeling (Stuker, 2008; Kanthak and Ney, 2003). These techniques, however, are usually language-dependent as linguistic knowledge of the target language has to be known in advance. In this paper, we will investigate a language-independent technique to improve current grapheme modeling.

Recently, we proposed a new method of estimating parameters of context-dependent models called *eigentriphone* acoustic modeling (Ko and Mak, 2011a; Ko and Mak, 2011b; Ko and Mak, 2012; Ko and Mak, accepted for publication). The main idea behind our method is the derivation of an eigenbasis over a set of triphones so that each triphone in the set can then be modeled as a distinct point in the space spanned by the basis vectors. The basis vectors are now called *eigentriphones* in our method. The eigentriphones represent the most important context-dependent characteristics among the given set of triphones. Since usually not many eigentriphones are required to represent the eigenspace, triphones with few training samples can be robustly estimated as a linear combination of the eigentriphones. From another point of view, context-dependent phonetic information is extracted from the more frequently occurring triphones in the triphone set in the form of basis vectors, which are shared with the less frequently occurring triphones in the set. The acoustic modeling of those triphones with limited amount of training data may then be thought of as an adaptation problem which is then solved by the eigenvoice approach (Kuhn et al., 2000). Moreover, compared with conventional tied-state triphone modeling, our new eigentriphone modeling method can eliminate the inevitable quantization error due to state tying — states tied together are not distinguishable. In our previous works, we investigated the use of model-based, state-based, and cluster-based eigentriphone acoustic modeling, and observed that cluster-based eigentriphone modeling consistently outperformed the conventional tied-state hidden Markov model (HMM) training method on TIMIT phoneme recognition and WSJ word recognition (Ko and Mak, accepted for publication).

Eigentriphone acoustic modeling is complementary to other acoustic modeling approaches such as SGMM that are based on tied-state HMM-GMMs. SGMM allows a compact representation of HMMs with a set of common basis Gaussians and state-dependent projection vectors. Since there are fewer parameters to estimate, model parameters of HMM-SGMM may be estimated more robustly for the same amount of training data. From another perspective, HMM-SGMMs may be trained with less data than HMM-GMMs. Whereas state are usually tied in HMM-SGMMs, eigentriphone modeling unties the tied states so that the states in the final models are, in general, all distinct. Eigentriphone modeling may take HMM-GMMs or HMM-SGMMs as its initial models, unties the states in each state cluster, derives an eigenbasis in each cluster, and re-estimates the mean vectors of each member state in the state cluster as a combination of the eigenvectors. As a result, eigentriphone modeling recovers the quantization errors in the initial tied-state HMM-GMMs or HMM-SGMMs so that the final models are more discriminative.

Although we call our method eigentriphone acoustic modeling, the proposed framework is actually very flexible and can be applied to any group of modeling units

Please cite this article in press as: Ko, T., Mak, B., Eigentrigraphemes for under-resourced languages, Speech Comm. (2013), http://dx.doi.org/10.1016/j.specom.2013.01.010

provided that they may be represented by vectors of the same dimension. In this paper, we would like to port the *cluster-based eigentriphone modeling* framework to a grapheme-based ASR system to estimate the parameters of trigrapheme acoustic models. The new method, which we call *cluster-based eigentrigrapheme acoustic modeling*, have the following favorable properties over other grapheme-based methods:

- Since it uses graphemes as the modeling units, it enjoys the same benefits that other grapheme-based modeling methods do. Most importantly, there is no need to create a phone set and a pronunciation dictionary. Thus, it is more favorable for building ASR system for under-resourced languages.
- Eigentrigrapheme modeling will also enjoy the same benefit like eigentriphone modeling: Many trigraphemes in under-resourced languages may have little training data; in the past, the problem is mainly solved by state tying, but eigentrigrapheme modeling allows reliable estimation of the infrequently occurring trigraphemes by careful state clustering and then projecting the member states of each cluster onto a low-dimensional subspace spanned by a small set of eigentrigraphemes of the cluster.
- No language-specific knowledge is required and the whole method is data-driven. It can be used to improve existing systems that are based on conventional tied-state trigrapheme HMM. In fact, one may implement our method as a post-processing procedure on conventional tied-state trigrapheme HMMs.
- If trigrapheme state clusters are created using graphemic decision tree, the decision tree may also be used to synthesize unseen trigraphemes in the test lexicon.
- States are generally not tied among the final trigrapheme models. Thus, the states are generally distinct from each other. We believe that the final trigrapheme models can be more discriminative than those trained on other methods.

The rest of this paper is organized as follows. In Section 2, we will describe the cluster-based eigentrigrapheme acoustic modeling method. That is followed by experimental evaluation in Section 3 and conclusions in Section 4.

2. Cluster-based eigentrigrapheme acoustic modeling

Fig. 1 shows an overview of the cluster-based eigentrigrapheme acoustic modeling method. The framework is very similar to the cluster-based eigentriphone modeling method (Ko and Mak, accepted for publication). All trigrapheme states are first represented by some supervectors and they are assumed to lie in a low dimensional space[1] spanned by a set of eigenvectors. In other words, each trigrapheme supervector is a linear combination of a small set of eigenvectors which are now called eigentrigraphemes. Clustering of the states can be done by a singleton decision tree, and the procedure is exactly the same as that of creating a conventional tied-state trigrapheme system.

Cluster-based eigentrigrapheme modeling consists of three major steps: (a) state clustering via a singleton decision tree, (b) derivation of the eigenbasis, and (c) estimation of eigentrigrapheme coefficients. They will be discussed in further details below.

2.1. Trigrapheme state clustering (or tying) by a singleton decision tree

One major difference between phone-based and grapheme-based acoustic modeling lies in the construction of the decision tree for tying hidden Markov model (HMM) states. In phone-based modeling, it is well-known that decision tree using phonetic questions (Young et al., 1994) can significantly improve speech recognition performance by striking a good balance between trainability and resolution of the acoustic models. However, it is not clear how the phonetic questions used in a phone-based system to tie triphone states can be ported to tie trigrapheme states in a grapheme-based system as the relation between the graphemes and their influence in the pronunciation of their neighboring graphemes is not well understood. Although the questions may be re-defined manually (Kanthak et al., 2002) or automatically, (Schukat-Talamazzini et al., 1993) investigated the performance of both methods in several languages and concluded that questions asking only the identity of the immediate neighboring grapheme, named as *singleton questions*, work at least as well as other types of questions. In this paper, decision tree[2] using singleton questions at each node is used to generate the conventional tied-state trigrapheme HMMs. In addition, the trigrapheme states that belong to the same tied state naturally form a state cluster on which our new cluster-based eigentrigrapheme modeling may be applied. In other words, the same singleton decision tree can be used to create the tied states for a conventional tied-state trigrapheme system as well as the state clusters for the construction of cluster-based eigentrigraphemes.[3]

[1] The dimension of the space is low when compared with the dimension of the trigrapheme state supervectors.

[2] The questions in the decision tree are generated from the grapheme set of the target language which is derived by scanning through the training data. Thus, the trees are language-dependent but our method is still language-independent.

[3] Although the state clusters of cluster-based eigentrigrapheme modeling and the tied states of conventional trigrapheme modeling both come from the nodes of the same decision tree, in general, they may not be exactly the same nodes. The optimal set of tied states or state clusters is determined using a separate set of development speech data.

Please cite this article in press as: Ko, T., Mak, B., Eigentrigraphemes for under-resourced languages, Speech Comm. (2013), http://dx.doi.org/10.1016/j.specom.2013.01.010

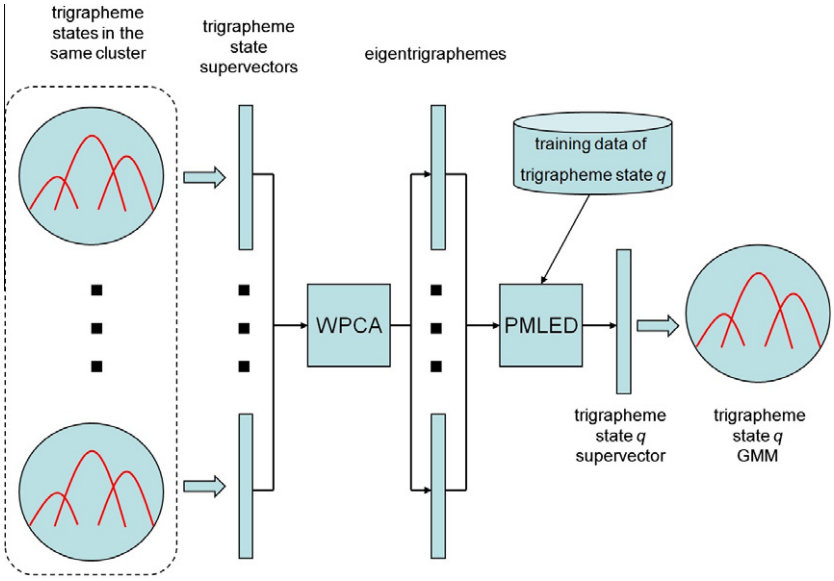

Fig. 1. The cluster-based eigentrigrapheme acoustic modeling method. (WPCA = weighted principal component analysis; PMLED = penalized maximum-likelihood eigen-decomposition).

2.2. Conventional tied-state trigrapheme HMM training

We adopt the standard procedure in HTK (Young et al., 2006) to create the conventional tied-state trigrapheme HMM system as follows.

STEP 1: Context-independent grapheme acoustic models are estimated from the training data. Each context-independent grapheme model is a 3-state strictly left-to-right HMM, and each state is represented by a single Gaussian.

STEP 2: Each context-independent grapheme HMM is then cloned to initialize *all* its context-dependent trigraphemes.

STEP 3: For each base grapheme, the transition probabilities of all its trigraphemes are tied together.

STEP 4: For each base grapheme, tie the corresponding HMM states of all its trigraphemes using a singleton decision tree. Thus, three singleton decision trees are built for each base grapheme. Once a set of trigrapheme states are tied together, they share the same set of Gaussian means, diagonal covariances, and mixture weights.

STEP 5: Synthesize the unseen trigraphemes by going through the singleton questions of the decision trees.

STEP 6: Grow the Gaussian mixtures of the models[4] with the training data until each tied state is represented by an M-component Gaussian mixture model (GMM) with diagonal covariance. In practice, the optimal value of M is determined by a separate set of development data.

[4] In this paper, 4 iterations of embedded Baum-Welch training was applied every time after the number of mixtures was grown.

2.3. Eigentrigrapheme acoustic modeling

Recall that each node in the state clustering decision tree has dual roles: it is treated as a tied state for tied-state HMM training, and as a state cluster for eigentrigrapheme modeling. To begin cluster-based eigentrigrapheme modeling, one first decides the tree nodes to be used as the state clusters. Then the state-clusters are treated as tied states, and conventional tied-state trigrapheme HMMs are created using the procedure described in Section 2.2. The resulting tied-state HMMs are used as the initial models for deriving the eigentrigraphemes of each state cluster.

2.3.1. Derivation of cluster-based eigentrigraphemes

The following procedure is repeated for each state cluster i, consisting of N_i member states.

STEP 7: Untie the Gaussian means of all the trigrapheme states in a state cluster with the exception of the unseen trigrapheme states. The means of the cluster GMM are then cloned to initialize *all* untied trigrapheme states in the cluster. Note that the Gaussian covariances and mixture weights of the states in the cluster are still tied together.

STEP 8: Re-estimate only the Gaussian means of trigrapheme states after cloning. Their Gaussian covariances and mixture weights remain unchanged as those of their state cluster GMM.

STEP 9: Create a trigrapheme state supervector $\mathbf{v}_{ip}$ for each trigrapheme state p in state cluster i by stacking up all its Gaussian mean vectors from its M-component GMM as below

$$\mathbf{v}_{ip} = \left[\boldsymbol{\mu}_{ip1}, \boldsymbol{\mu}_{ip2}, \ldots, \boldsymbol{\mu}_{ipM}\right], \quad (1)$$

where $\boldsymbol{\mu}_{ipm}, m = 1, 2, \ldots, M$ is the mean vector of the mth Gaussian component[5]. Similarly, a state cluster supervector $\mathbf{m}_i$ is created from the GMM of state cluster i.

STEP 10: Collect the trigrapheme state supervectors $\{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{iN_i}\}$ as well as the state cluster supervector $\mathbf{m}_i$ of cluster i, and derive an eigenbasis from their correlation matrix using *weighted principal component analysis* (WPCA). The correlation matrix is computed as follows:

$$\frac{1}{n_i}\sum_p n_{ip}(\hat{\mathbf{v}}_{ip} - \hat{\mathbf{m}}_i)(\hat{\mathbf{v}}_{ip} - \hat{\mathbf{m}}_i)', \quad (2)$$

where $\hat{\mathbf{v}}_{ip}$ and $\hat{\mathbf{m}}_i$ are the standardized version of $\mathbf{v}_{ip}$ and $\mathbf{m}_i$ that are created by normalizing them with the diagonal covariance matrix; n_{ip} is the frame count of the trigrapheme state p in cluster i, and $n_i = \sum_p n_{ip}$ is the total frame counts of state cluster i. A comparison of using the standard PCA and WPCA in deriving the eigenbasis (Ko and Mak, 2012) shows that WPCA is more effective because the eigenvalues fall faster so that fewer eigenvectors are sufficient for representing the eigentrigrapheme space. Therefore, WPCA is used throughout this paper.

STEP 11: Arrange the eigenvectors $\{\hat{\mathbf{e}}_{ik}, k = 1, 2, \ldots, N_i\}$ in descending order of their eigenvalues λ_{ik}, and pick the top K_i (where $K_i \leqslant N_i$) eigenvectors to represent the eigenspace of state cluster i. These K_i eigenvectors are now called *eigentrigraphemes* of state cluster i. Note that different state clusters may have a different number of eigentrigraphemes.

2.3.2. Estimation of the eigentrigrapheme coefficients

After the derivation of the eigentrigraphemes, the supervector $\mathbf{v}_{ip}$ of any trigrapheme state p in cluster i is assumed to lie in the space spanned by the K_i eigentrigraphemes. Thus, we have

$$\mathbf{v}_{ip} = \mathbf{m}_i + \sum_{k=1}^{K_i} w_{ipk}\mathbf{e}_{ik}, \quad (3)$$

where $\mathbf{e}_{ik}, k = 1, 2, \ldots, K_i$ is the rescaled version of the standardized eigenvector $\hat{\mathbf{e}}_{ik}$; $\mathbf{w}_{ip} = [w_{ip1}, w_{ip2}, \ldots, w_{ipK_i}]$ is the eigentrigrapheme coefficient vector of trigrapheme state p in the trigrapheme state space of cluster i.

The eigentrigrapheme coefficient vector $\mathbf{w}_{ip}$ is estimated by maximizing the following penalized likelihood objective function $Q(\mathbf{w}_{ip})$:

$$Q(\mathbf{w}_{ip}) = L(\mathbf{w}_{ip}) - \beta R(\mathbf{w}_{ip}), \quad (4)$$

where β is the regularization parameter that balances the dynamic ranges of the regularizer $R(\cdot)$ and the likelihood term $L(\cdot)$. $L(\mathbf{w}_{ip})$ is the likelihood of the training data of trigrapheme state p in cluster i, and is given by

$$L(\mathbf{w}_{ip}) = \text{constant} - \sum_{m,t}\gamma_{ipm}(t)(\mathbf{x}_t - \boldsymbol{\mu}_{ipm}(\mathbf{w}_{ip}))'C_{ipm}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{ipm}(\mathbf{w}_{ip})) \quad (5)$$

where C_{ipm} and $\gamma_{ipm}(t)$ are the covariance and occupation probability of the mth Gaussian of trigrapheme state p in cluster i given observation $\mathbf{x}_t$.

To avoid the estimation of $\mathbf{w}_{ip}$ from over-fitting, the regularizer $R(\cdot)$ is introduced in the objective function. From Eq. (3), one can see the $|\mathbf{w}_{ip}|$ is the Euclidean distance of the trigrapheme state supervector $\mathbf{v}_{ip}$ from its state cluster supervector $\mathbf{m}_i$ in the trigrapheme state space. In Ko and Mak (2011), the following regularizer

$$R(\mathbf{w}_{ip}) = \sum_{k=1}^{K_i}\frac{w_{ipk}^2}{\lambda_{ik}} \quad (6)$$

is found effective. It has the following properties:

- The squared coefficient of each eigentrigrapheme, w_{ipk}, is inversely scaled by its eigenvalue so that a less informative eigentrigrapheme (with smaller eigenvalues) will have less influence on the "adapted" trigrapheme model. On the other hand, the trigrapheme can freely move along the more informative eigentrigraphemes (with larger eigenvalues) in the trigrapheme state space.
- Since the chosen regularizer automatically deemphasizes the less informative eigentrigraphemes, we may avoid making a hard decision on the number of eigentrigraphemes K_i for each state cluster. Instead, we may simply use *all* eigentrigraphemes (by setting $K_i = N_i$) for every state cluster, eliminating the need to tune the value of K_i.
- When a trigrapheme state has a lot of training data, the likelihood term will become dominant in the objective function $Q(\mathbf{w}_{ip})$. As a result, the "adapted" model should converge to its Baum–Welch training estimate.
- On the other hand, for a trigrapheme state with limited amount of training data, the regularizer becomes dominant and it will try to make $|\mathbf{w}_{ip}|$ as small as possible. Thus, the "adapted" model will fallback to its state cluster GMM.

The above estimation method of the trigrapheme coefficients can be viewed as a penalized version of the *maximum-likelihood eigen-decomposition* (MLED) in eigenvoice adaptation (Kuhn et al., 2000), and it is called *penalized maximum-likelihood eigen-decomposition* (PMLED) in Fig. 1.

Differentiating the optimization function $Q(\mathbf{w}_{ip})$ of Eq. (4) w.r.t. each eigentrigrapheme coefficient w_{ipk}, and setting each derivative to zero, we have,

[5] Since the mixture weights are still tied among the trigrapheme states in a state cluster, the M Gaussian components in each state can be consistently ordered across all the member states in the cluster to create their supervectors.

$$\sum_{n=1}^{K_i} A_{ipkn} w_{ipn} + \beta \frac{w_{ipk}}{\lambda_{ik}} = B_{ipk}, \quad \forall k = 1, 2, \ldots K_i \tag{7}$$

where

$$A_{ipkn} = \sum_m \mathbf{e}'_{ikm} C^{-1}_{ipm} \mathbf{e}_{inm} \left(\sum_t \gamma_{ipm}(t) \right)$$

$$B_{ipk} = \sum_m \mathbf{e}'_{ikm} C^{-1}_{ipm} \left(\sum_t \gamma_{ipm}(t)(\mathbf{x}_t - \mathbf{m}_{im}) \right).$$

The eigentrigrapheme coefficients may be easily found by solving the system of K_i linear equations represented by Eq. (7), and the Gaussian mean of the mth mixture of trigrapheme state p in cluster i can be obtained from $\mathbf{v}_{ip}$ as

$$\boldsymbol{\mu}_{ipm} = \mathbf{m}_{im} + \sum_{k=1}^{K_i} w_{ipk} \mathbf{e}_{ikm}. \tag{8}$$

The eigentrigrapheme modeling procedure stops if either the estimation of eigentrigrapheme coefficients converges or the recognition accuracy of the trained models is maximum on a development data set. Otherwise, the training data are re-aligned with the current models, and the derivation of eigentrigraphemes and the estimation of their coefficients are repeated.

3. Experimental evaluation

The effectiveness of our newly proposed eigentrigrapheme acoustic modeling method is evaluated on four under-resourced languages of South Africa with the assumption that no phonetic dictionaries are available. Since graphemes are the basic modeling units in grapheme-based modeling, word recognition accuracy is the main metric for the evaluation. Nonetheless, triphone-based systems were also built with the use of semi-automatically generated phonetic dictionaries so as to benchmark the results of our eigentrigrapheme results (where no dictionaries are used).

3.1. The Lwazi speech corpus

The Lwazi project was set up to develop a telephone-based speech-driven information system to take advantage of the more and more popular use of telephones in South Africa nowadays. As part of the project, the Lwazi ASR corpus(2009) was collected to provide the necessary speech and language resources in building ASR systems for all eleven official languages of South Africa.

The corpus was collected from approximately 200 speakers per language who are all first language speakers. Each speaker produced approximately 30 utterances, in which 16 of them are phonetically balanced read speech and the remainders are elicited short words such as answers to open questions, answers to yes/no questions, spelt words, dates, and numbers. All the data were recorded over a telephone channel and were transcribed only in words. Background noises, speaker noises, and partial words are marked in the orthographic transcriptions.

The Lwazi project also created a 5,000-word pronunciation dictionary for each language (Davel and Martirosian, 2009). These dictionaries cover the most common words in the language but not all the words appearing in the corpus. Thus, for the phone-based experiments, the *DictionaryMaker* (Tempest and Davel, 2009) software was used to generate dictionary entries for the words that are not covered by the Lwazi dictionaries. The given Lwazi dictionaries were used as the seed dictionaries[6] for DictionaryMaker to extract grapheme-to-phoneme conversion rules which were then applied to generate the phonetic transcriptions for the uncovered words for each language. The pronunciations suggested by DictionaryMaker were directly used without any modification.

Among the eleven South African official languages, four are chosen for this investigation. We looked at their ranks according to three different criteria:

- the human language technology (HLT) index (Sharma-Grover et al., 2010): the index indicates the total quantity of HLT activity for each language. Higher the index is, greater HLT development has been done.
- the phone recognition accuracy (van Heerden et al., 2009): higher phone accuracy means a higher rank for the language.
- the amount of training data available (van Heerden et al., 2009): language with more training data will be given a higher rank.

Finally, the following four languages are chosen because they have a good mix of phone accuracies and HLT activities as shown in Table 1:

Afrikaans: Afrikaans is a Low Franconian, West Germanic language, originated from Dutch (van Huyssteen and Pilon, 2009). It has about 6 million native speakers and is the third largest language in South Africa. It is also spoken in South Africa's neighboring countries like Namibia, Botswana and Zimbabwe. It has relatively more resources (Roux et al., 2004), and more ASR related works (de Wet et al., 2011; Kamper and Niesler, 2011) have been done on it than other languages of South Africa. It is interesting to see that although Afrikaans has the least amount of training data in the corpus, its phone recognition result is quite good among the eleven South African languages.

South African (SA) English: SA English is the de facto South Africa's lingua franca. It is spoken by about 3.6 million people in South Africa. SA English

[6] For Afrikaans, the dictionary available at http://sourceforge.net/projects/rcrl/files/AfrPronDict/ was used together with the Lwazi dictionary as the seed dictionary.

Table 1
Ranks of the four chosen South African languages in three aspects: their human language technology (HLT) indices, phone recognition accuracies, and amount of training data in the Lwazi corpus. (Smaller value implies a higher rank.).

Language	HLT rank (Sharma-Grover et al., 2010)	Phone recognition (van Heerden et al., 2009)	Amount of data (van Heerden et al., 2009)
Afrikaans	1	5	11
SA English	2	11	10
Sesotho	7	7	7
siSwati	9	3	1

is evolved from British English but is highly influenced by Afrikaans and the other languages of the country.

Sesotho: Sesotho is a Southern Bantu language, closely related to other languages in the Sotho-Tswana language group. It has about 3.5 million native speakers and is the seventh largest language in South Africa.

siSwati: siSwati is also a Southern Bantu language, closely related to the Nguni language group. It has about 1 million native speakers and is the ninth largest language in South Africa.

Since the corpus does not define the training, development, and test set for each language, we did the partitions ourselves. The data sets used in our experiments are summarized in Table 2. It is interesting to see that languages with more training data (in terms of duration) have a higher percentage of out-of-vocabulary words in their test set.

3.2. Feature extraction and common experimental settings

The first 13 PLP coefficients, including c0, and their first and second order derivatives were used. These 39-dimensional feature vectors were extracted at every 10 ms over a window of 25 ms. Speaker-based cepstral mean subtraction and variance normalization were performed.

The grapheme set and phone set of each language are the same as the ones defined in the Lwazi dictionaries.

Table 2
Information of the data sets of four South African languages used in this investigation. (OOV is *out-of-vocabulary*).

Data set	#Speakers	#Utt.	Dur.(hr)	Vocab	OOV%
Afrikaans training	160	4784	3.37	1513	0.00
Afrikaans dev.	20	600	—	870	0.89
Afrikaans test	20	599	—	876	0.97
SA English training	156	4665	3.98	1988	0.00
SA English dev.	20	581	—	1104	1.10
SA English test	20	597	—	1169	1.68
Sesotho training	162	4826	5.70	2360	0.00
Sesotho dev.	20	600	—	1096	1.86
Sesotho test	20	601	—	1089	2.29
siSwati training	156	4643	8.38	4645	0.00
siSwati dev.	20	599	—	1889	6.14
siSwati test	20	596	—	1851	4.53

For all systems described below, transition probabilities of all triphones/trigraphemes of the same base phone/grapheme were tied together. Each triphone/trigrapheme model was a strictly left-to-right 3-state continuous-density hidden Markov model (CDHMM) with a Gaussian mixture density of at most $M = 16$ components per state. In addition, there were a 1-state short pause model and a 3-state silence model whose middle state was tied with the short pause state. Recognition was performed using the HTK toolkit (Young et al., 2006) with a beam search threshold of 350. Only the annotated text data in the training set were used to train the corresponding language models. Both phone trigram language models and word bigram language models were estimated for the four languages except Sesotho, for which only phone bigrams could be reliably trained. Perplexities of the various language models on the development data and test data are shown in Table 3.

All system parameters such as the grammar factor, insertion penalty, regularization parameter β, number of GMM components M, number of tied states or state clusters, and so forth were optimized using the respective development data.

3.3. Phone and word recognition using triphone-based HMMs

We first establish the triphone-based ASR benchmarks against which the trigrapheme-based models can be checked. Both conventional tied-state triphone HMM modeling and our new cluster-based eigentriphone modeling were tried for the four under-resourced languages of South Africa. The number of base phones, the number of cross-word triphones in the training set, the optimal number of tied states in conventional HMM training, and the optimal number of state clusters in eigentriphone modeling for each language are summarized in Table 4.

3.3.1. Phone recognition results

Although word recognition accuracy will be the eventual evaluation metric for grapheme modeling, we also would like to report the phone recognition baselines of our triphone models for the sake of completeness. Phone recognition was performed on each of the four languages using no

Table 3
Perplexities of phone and word language models of the four South African languages.

Language	Data set	Phone perplexity	Word bigram perplexity
Afrikaans	Dev.	7.37 (trigram)	12.4
	Test	7.33 (trigram)	11.18
SA English	Dev.	7.50 (trigram)	13.28
	Test	7.76 (trigram)	11.18
Sesotho	Dev.	10.43 (bigram)	19.60
	Test	10.29 (bigram)	19.69
siSwati	Dev.	7.60 (trigram)	12.27
	Test	7.50 (trigram)	10.94

Please cite this article in press as: Ko, T., Mak, B., Eigentrigraphemes for under-resourced languages, Speech Comm. (2013), http://dx.doi.org/10.1016/j.specom.2013.01.010

Table 4
Some system parameters of triphone modeling in the four South African languages.

Language	#Base phones	#Cross-word triphones	#Tied states in conventional models	#State clusters in eigentriphone models
Afrikaans	37	5203	617	332
SA English	44	7167	988	362
Sesotho	41	4061	741	624
siSwati	40	5140	339	250

Table 6
Word recognition accuracy (%) of four South African languages.

Language	Tied-state triphone		Cluster-based eigen-trigrapheme/-triphone	
	Trigrapheme	Triphone	Trigrapheme	Triphone
Afrikaans	89.39	89.73	89.87	90.73
SA English	78.30	83.12	79.57	83.72
Sesotho	75.67	75.57	76.35	76.77
siSwati	80.04	79.79	80.67	80.29

or flat LM as well as using its respective bigram/trigram LM. The results are given in Table 5. The following observations can be made:

- Our phone recognition results with flat LMs are quite different from those reported in van Heerden et al. (2009). There may be a few reasons:
  - To our knowledge, the Lwazi corpus has been evolving, and the corpus we obtained earlier this year is different from the older version used in van Heerden et al. (2009).
  - Since there are no official test sets in the corpus, it is hard to compare recognition performance from different research groups.
  - Since the data are not manually labeled by professional transcribers, there is no ground truth which the results from different research groups can compare with.

  Thus, it may not be meaningful to compare our phone recognition results with others. We believe it is good enough to see that our results are in the same ballpark as the others.
- SA English has substantially lower phone recognition accuracy: it is lower than that of the other three languages by more than 10% absolute. Although SA English has a few more phones in its phonetic inventory than the other languages, and significantly more cross-word triphones to model (see Table 4), its phone trigram perplexity is actually similar to Afrikaans and siSwati. (Only bigram language model can be reliably estimated for Sesotho, and its value is expected to be higher than the phone trigram perplexity of the other three languages.) It means that the phone trigrams (as well as triphones) of SA English are more unevenly distributed in the training corpus.The lower phone recognition accuracy of SA English may be simply due to its larger inventory of phones and triphones, making discrimination among them more difficult. Another plausible reason is that SA English is now the de facto lingua franca of South Africa. It is usually the language of choice for communication among people from different regions and ethnic groups of the country including immigrants from China and India. As a consequence, there are more allophonic variations in SA English, making it harder to recognize.
- The training speech data are not phonetically labelled by human transcribers. Instead, their phonetic transcriptions are generated semi-automatically by grapheme-to-phoneme conversion together with a small bootstrapping dictionary. From the big improvement of recognition performance when phone language models were used (vs. when no language models were used), we may conclude that phone language models trained from the generated phonetic transcriptions are good enough to improve phone recognition significantly.
- Triphone models estimated by our new cluster-based eigentriphone modeling method outperform triphone models estimated by conventional tied-state HMM training by an average of 6.19% relative over the four languages.

3.3.2. Word recognition results

The word recognition performance of the triphone-based systems are shown in Table 6. We can see that

- With no surprise, Sesotho, having the highest LM perplexity (see Table 3), has the lowest recognition accuracy.

Table 5
Phone recognition accuracy (%) of four South African languages. († The benchmark results in van Heerden et al. (2009) used an older version of the Lwazi corpus and how the corpus were partitioned into training, development, and test sets is unknown.)

Language	Benchmark (van Heerden et al., 2009)†	Tied-state triphone		Cluster-based eigentriphone	
	Flat LM	Flat LM	N-gram LM	Flat LM	N-gram LM
Afrikaans	63.14	59.07	69.73 (trigram)	62.23	72.32 (trigram)
SA English	54.26	45.48	56.58 (trigram)	46.03	57.84 (trigram)
Sesotho	54.79	62.36	67.06 (bigram)	64.08	68.35 (bigram)
siSwati	64.46	64.76	71.45 (trigram)	68.19	74.13 (trigram)

Please cite this article in press as: Ko, T., Mak, B., Eigentrigraphemes for under-resourced languages, Speech Comm. (2013), http://dx.doi.org/10.1016/j.specom.2013.01.010

Table 7
Some system parameters used in trigrapheme modeling of the four South African languages. (The numbers of possible base graphemes are 43, 26, 27, 26 for the four languages but not all of them are seen in the corpus.)

Language	#Seen base graphemes	#Cross-word trigraphemes	#Tied states in conventional models	#State clusters in eigentrigrapheme models
Afrikaans	31	3458	728	332
SA English	26	4125	1630	547
Sesotho	25	3072	543	543
siSwati	25	3826	392	255

- For the other three languages, namely Afrikaans, SA English, and siSwati, which all have similar word bigram perplexity, their word recognition performance is well correlated with their vocabulary size and OOV figure. Afrikaans has the best word recognition accuracy, and yet there are only 1513 words in its vocabulary with 0.97% OOV. On the other hand, siSwati has the worst performance, and its vocabulary size is 4,645 with 4.53% OOV, which are 3–4 times of those of Afrikaans (see Table 4).
- Although SA English has the poorest phone recognition accuracy, its word recognition performance is second among the four languages. It does not only show the limitation of using phone recognition accuracy to predict word recognition performance, but also the effectiveness of a good n-gram language model for word recognition.
- Cluster-based eigentriphone modeling outperforms conventional tied-state HMM training by an average of 5.17% relative over the four languages.

3.4. Word recognition using trigrapheme-based HMMs

Similar acoustic models were developed using trigraphemes; there is no need for a phonetic dictionary in the process. The number of base graphemes actually observed in the corpus, the number of cross-word trigraphemes in the training set, the optimal number of tied states in conventional HMM training, and the optimal number of state clusters in eigentrigrapheme modeling for each language are summarized in Table 7. The word recognition results of the various trigrapheme-based systems are shown in Table 6 together with the results from the corresponding triphone-based systems so that they can be easily compared.

Besides the observations that are mentioned in triphone-based systems in Section 3.3.2, the following additional observations are well noted.

- Except for SA English, our trigrapheme-based systems performs basically the same as their triphone-based counterparts even without the knowledge of a phonetic dictionary. In fact, trigrapheme-based systems even outperform their triphone-based counterpart in siSwati though insignificantly. The results suggest that there is a consistent mapping between the pronunciation of Afrikaans, Sesotho, and siSwati and their graphemes.
- Trigrapheme-based systems perform much worse than triphone-based systems in SA English. This is expected. Similar results have been reported for English (Stuker, 2009). Besides the reason mentioned in the Introduction Section that the pronunciation of English has developed away from its written form over time, the particularly great allophonic variations in SA English (which is also reflected in its phone recognition accuracy) further compromise the word recognition effort.
- Once again, our new cluster-based eigentrigrapheme modeling consistently performs better than conventional tied-state trigrapheme HMM training. If has an average gain of 4.08% relative over the four languages.

4. Conclusions

Most state-of-the-art automatic speech recognition (ASR) systems are developed using phonetic acoustic models. However, for many developing or under-developed countries in the world, the adoption of human language technologies has been dragged down by the lack of speech and language resources, which are usually costly and take a lot of human expertise to acquire. Graphemic acoustic modeling mitigates the problem as it does not require a phonetic dictionary. In this paper, we port a new acoustic modeling method called cluster-based eigentriphone modeling which has been shown to outperform the conventional tied-state triphone HMM training in phone-based ASR systems to grapheme-based ASR for under-resourced languages. We call the new method *cluster-based eigentrigrapheme acoustic modeling*.

For four under-resourced languages of South Africa (SA), namely, Afrikaans, SA English, Sesotho, and siSwati, it is shown that in terms of word recognition performance, trigrapheme-based ASR is as good as triphone-based ASR with the exception of SA English. The worse performance of trigrapheme-based ASR on SA English is not unexpected, since SA English is a variation of British English, and it is well known that grapheme-based ASR does not perform well on the latter. In particular, trigrapheme acoustic models trained by our new eigentrigrapheme mod-

Please cite this article in press as: Ko, T., Mak, B., Eigentrigraphemes for under-resourced languages, Speech Comm. (2013), http://dx.doi.org/10.1016/j.specom.2013.01.010

eling method consistently outperforms the trigrapheme models trained by conventional tied-state HMM training, achieving a relative reduction in the word error rates of the four SA languages by an average of 4.08%. Trigrapheme HMM states trained by the eigentrigrapheme modeling method are distinct from each other — the quantization error among the member states of a tied state in conventional HMM is avoided — and should be more discriminative.

In the future, we would like to investigate other cluster definitions for under-resourced language ASR, especially when the amount of acoustic training data is even smaller. The effect of discriminative training on the inherently distinctive models produced by eigentrigrapheme modeling will be studied as well.

Acknowledgement

This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant numbers FSGRF12EG31, FSGRF13EG20, and SRFI11EG15.

References

Ogbureke, K.U., Carson-Berndsen, J., 2010. Framework for cross-language automatic phonetic segmentation. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and, Signal Processing.

Le, V.-B., Besacier, L., 2009. Automatic speech recognition for under-resourced languages: application to Vietnamese language. IEEE Trans. Audio Speech Lang. Process. 17, 1471–1482.

Kohler, J., 1996. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In: Proc. Internat. Conf. on Spoken Language Processing.

Beyerlein, P. et al., 1999. Towards language independent acoustic modeling. In: Proc. IEEE Automatic Speech Recognition and Understanding, Workshop.

Sooful, J.J., Botha, E.C., 2001. An acoustic distance measure for automatic cross-language phoneme mapping. In: Proc. Pattern Recognition Association of South Africa.

Povey, D. et al., 2010. Subspace Gaussian mixture models for speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafit, M., Povey, D., Rastrow, A., Rose, R., Thomas, S., 2010. Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

Lu, L., Ghoshal, A., Renals, S., 2011. Regularized subspace gaussian mixture models for cross-lingual speech recognition. In: 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 365–370.

Meng, H., Hunnicutt, S., Seneff, S., Zue, V., 1996. Reversible letter-to-sound generation based on parsing word morphology. Speech Communications 18, 47–63.

Bellegarda, J.R., 2003. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

Davel, M., Barnard, E., 2004. The efficient creation of pronunciation dictionaries: human factors in bootstrapping. In: Proc. Interspeech.

Andersen, O., Kuhn, R., Lazarides, A., Dalsgaard, P., Haas, J., Noth, E., 1996. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion. In: Proc. Internat. Conf. on Spoken Language Processing.

Schukat-Talamazzini, E.G., Niemann, H., Eckert, W., Kuhn, T., Rieck, S., 1993. Automatic speech recognition without phonemes. In: Proc. European Conf. on Speech Communication and Technology.

Kanthak, S., Ney, H., 2002. Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

Charoenpornsawat, S.H.P., Schultz, T., 2006. Thai grapheme-based speech recognition. In: Proc. Human Language Technology Conf. of the North American Chapter of the ACL.

Daniels, P.T., Bright, W. (Eds.), 1996. The World's Writing Systems. Oxford University Press, 198 Madison Avenue, New York, NY 10016, USA.

Stuker, S., 2009. Acoustic modeling for under-resourced languages. Ph.D. thesis, Universität Fridericiana zu Karlsruhe (TH).

Stuker, S., 2008. Modified polyphone decision tree specialization for porting multilingual grapheme based ASR systems to new languages. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

Kanthak, S., Ney, H., 2003. Multilingual acoustic modeling using graphemes. In: Proc. European Conf. on Speech Communication and Technology.

Ko, T., Mak, B., 2011a. Eigentriphones: a basis for context-dependent acoustic modeling. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

Ko, T., Mak, B., 2011b. A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization. In: Proc Interspeech.

Ko, T., Mak, B., 2012. Derivation of eigentriphones by weighted principal component analysis. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing.

Ko, T., Mak, B., accepted for publication. Eigentriphones for context-dependent acoustic modeling. IEEE Trans. Audio Speech Lang. Process, http://dx.doi.org/10.1109/TASL.2013.2248722.

Kuhn, R., Junqua, J.-C., Nguyen, P., Niedzielski, N., 2000. Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process. 8, 695–707.

Young, S.J., Odell, J.J., Woodland, P.C., 1994. Tree-based state tying for high accuracy acoustic modelling. In: Proc. Workshop on Human Language Technology.

Young, S. et al., 2006. The HTK Book (Version 3.4), University of Cambridge.

Meraka-Institute, Lwazi ASR corpus. <http://www.meraka.org.za/lwazi>.

Davel, M., Martirosian, O., 2009. Pronunciation dictionary development in resource-scarce environments, in: Proc. Interspeech.

Tempest, M., Davel, M., 2009. Dictionary Maker 2.16 user manual. <http://dictionarymaker.sourceforge.net>.

Sharma-Grover, A., van Huyssteen, G.B., Pretorius, M.W., 2010. An HLT profile of the official South African languages. In: Proc. Second Workshop on African Language Technology (AfLaT 2010).

van Heerden, C., Barnard, E., Davel, M., 2009. Basic speech recognition for spoken dialogues. In: Proc. Interspeech.

van Huyssteen, G.B., Pilon, S., 2009. Rule-based conversion of closely-related languages: a Dutch-to-Afrikaans convertor. In: Proc. 20th Annual Symposium of the Pattern Recognition Association of South Africa.

Roux, J.C., Louw, P.H., Niesler, T.R., 2004. The African speech technology project: an assessment. In: Proc. LREC.

de Wet, F., de Waal, A., van Huyssteen, G.B., 2011. Developing a broadband automatic speech recognition system for Afrikaans. In: Proc. Interspeech.

Kamper, H., Niesler, T., 2011. Multi-accent speech recognition of Afrikaans, black and white varieties of South African English. In: Proc. Interspeech.