Engineering Applications of Artificial Intelligence 40 (2015) 95–102

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Unsupervised rapid speaker adaptation based on selective eigenvoice merging for user-specific voice interaction

Dong-Jin Choi [a,1], Jeong-Sik Park [b,*], Yung-Hwan Oh [a]

[a] Department of Computer Science, Korea Advanced Institute of Science and Technology, 291 Daehak-Ro, Daejeon, Republic of Korea
[b] Department of Information and Communication Engineering, Yeungnam University, 280 Daehak-Ro, Gyeongsan, Republic of Korea

ARTICLE INFO

Article history:
Received 23 September 2014
Received in revised form
8 January 2015
Accepted 16 January 2015
Available online 13 February 2015

Keywords:
Speaker adaptation
Eigenvoice
Maximum Likelihood Linear Regression
Maximum A Posteriori
Selective eigenvoice merging
Speech recognition

ABSTRACT

Speaker adaptation transforms the standard speaker-independent acoustic models into an adapted model relevant to the user (called the target speaker) in order to provide reliable speech recognition performance. Although several conventional adaptation techniques, such as Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP), have been successfully applied to speech recognition tasks, they demonstrate great dependency on the amount of adaptation data. However, the eigenvoice-based adaptation technique is known to provide reliable performance regardless of the amount of data, even for a very small amount. In this study, we propose an efficient eigenvoice adaptation approach to construct more reliable adapted models. The proposed approach merges eigenvoice sets for possible eigenvoice combinations, and then selects optimal eigenvoice sets that are most relevant to the target speaker. For this task, we propose an efficient unsupervised eigenvoice selection method as well as a rapid merging technique. On speech recognition experiments using the Defense Advanced Research Projects Agency's Resource Management corpus, the proposed approach exhibited superior performance, compared to conventional methods, in both recognition accuracy and time complexity.

1. Introduction

As various types of smart devices, such as smartphones, smart TVs, and even the forthcoming humanoid robots, have been introduced, the necessity for convenient and user-friendly ways of interaction between device and user has also increased. Among many useful interactive interfaces, voice is the most convenient and natural while conveying useful and intuitive non-verbal information as well as verbal information (Fong et al., 2003). Since the advent of voice-based search or voice-driven communication through mobile applications, human–machine interaction via voice interface is not a new idea at all. In addition to human speech recognition, people now expect to experience more natural and effective voice interaction with machines.

Even though significant advancements in speech recognition technology motivated smart devices's ability to correctly recognize continuous human speech, it is still a challenge to recognize the natural or conversational voice in adverse environments (Ting et al., 2013; Amrouche et al., 2010). In particular, most speech recognition systems tend to demonstrate different performance among speakers, only working well with speakers adhering to the characteristics of the acoustic model (e.g. Hidden Markov Model (HMM)) constructed for the system. Such a limitation is a more dominant feature in speech emotion recognition tasks, in which acoustic models may rarely handle human's different ways of expressing emotional states, thus producing considerable performance variations among speakers (Park et al., 2009).

This tendency for speaker-dependent recognition performance is mainly induced by different characteristics between a system user and a set of speakers involved in the construction of acoustic models. Such an acoustic model is called the speaker-independent (SI) model, which is constructed with an amount of speech data obtained from a specific group of speakers, irrespective of the system users. Another type of acoustic model is the speaker-dependent (SD) model, which is built with a sufficient amount of speech data collected only from one user. SD models preserve the user's acoustic characteristics, thus providing the best recognition performance. They are ideal for real-world devices, but not practical because of the difficulty of collecting a sufficient amount of speech data from individual users.

Speaker adaptation techniques were introduced to overcome the drawbacks of the two types of acoustic model (Shinoda, 2011). Speaker adaptation transforms the standard SI models using a small amount of speech data obtained from the system user (also called a target speaker), as illustrated in Fig. 1. The transformed

* Corresponding author.
E-mail address: parkjs@yu.ac.kr (J.-S. Park).
[1] Current address: Daum Communications, 242, Cheomdan-ro, Jeju-si, Jeju-do, Republic of Korea.

http://dx.doi.org/10.1016/j.engappai.2015.01.010
0952-1976/

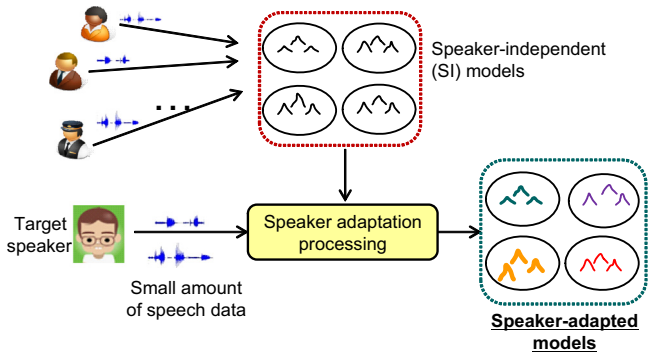

Fig. 1. General concept of speaker adaptation.

models have characteristics adapted to the target speaker, and convey acoustic characteristics similar to the speaker, thus providing the user with better performance compared to the SI model. Although many adaptation techniques have been successfully applied to speech recognition tasks, they still have weaknesses in voice interaction with real-world devices that are generally accessed by various anonymous users. In this paper, we propose an efficient speaker adaptation technique for handling voice interaction between a device and unspecified users.

This paper is organized as follows. Section 2 introduces the conventional speaker adaptation techniques and their drawbacks in real-world applications. Section 3 describes the proposed method in detail. Section 4 demonstrates and discusses the experimental results. Finally, conclusions are presented in Section 5.

2. Conventional speaker adaptation approaches and their efficiency in real-world applications

2.1. Conventional adaptation approaches

Many speaker adaptation techniques have been studied to reduce mismatches of speaker characteristics between training data and testing data (Shinoda, 2011). In addition to automatic speech recognition, recent studies have attempted to apply this technique to other applications to solve similar mismatch problems (Kim et al., 2012; Li and Dong, 2013; Mohammadi et al., 2014).

The adaptation techniques are generally classified into three main categories (Shinoda, 2011; Woodland, 2001): parameter transformation-based approaches using Maximum Likelihood Linear Regression (MLLR) (Gales, 1998), parameter re-estimation-based approaches such as Maximum A Posteriori (MAP) (Gauvain and Lee, 1994), and speaker clustering-based approaches like the eigenvoice technique (Kuhn et al., 1998).

The parameter transformation-based approaches assume that a set of model parameters, such as means and variances in the Gaussian Mixture Model (GMM), can be transformed by a type of transformation matrix. The basic procedure is to calculate one matrix or several transformation matrices from adaptation data collected from a target speaker. Parameter re-estimation-based approaches, also called the Bayesian adaptation, re-estimate individual model parameters, using a priori knowledge (usually SI model parameters) on the basis of a Bayesian framework. MLLR and MAP are known as the standard successful adaptation techniques for each of the two approaches. But, they operate well only when sufficient adaptation data are available (Kuhn et al., 2000; Goronzy and Kompe, 1999; Wang, 2003). The eigenvoice-based speaker adaptation approach was introduced to overcome this drawback of MLLR and MAP. The fundamental concept of this approach is that the acoustic characteristics of a target speaker can be defined by a linear combination of a small number of parameters obtained from existing SD models. The representative parameters are called eigenvoices. The main objective is to define the eigenvoices representing relevant characteristics of the target speaker.

The conventional adaptation techniques can be handled using two types of learning approach: supervised and unsupervised (Matsui and Furui, 1998; Wallace et al., 2009). In supervised adaptation, prior knowledge of adaptation data, such as speaker information or manually defined labels, is required. On the other hand, unsupervised approaches determine the labels of adaptation data by recognizing them automatically and do not require speaker-specific information. For supervised adaptation, every target speaker is encouraged to record predetermined words or sentences in the system to obtain adaptation data, whereas an unsupervised adaptation system allows the user to speak any words or sentences freely.

2.2. Efficiency of the conventional approaches in real-world application to voice interaction

Speaker adaption techniques are capable of adjusting the voice interaction system adopted in smart devices toward a specific user. Nevertheless, the conventional techniques have weaknesses in real-world application because most of them concentrate on investigating efficiency for either a small amount or a large amount of adaptation data without the consideration of general user cases.

In general environments, the period of interaction with a device may be different for each user. For instance, a smart device like a service robot that confronts strangers in a public space can have a short-term or a long-term conversation with a user. This tendency concludes that the voice interaction system may obtain different amounts of adaptation data according to the target speaker. Therefore, a desirable speaker adaptation technique should always provide reliable performance regardless of the amount of adaptation data. As addressed in Section 2.1, the performance of MLLR and MAP relies highly on the amount of adaptation data. In general, MLLR adaptation demonstrates better performance with a small amount of data. However, the adaptation process becomes saturated at some point—its performance rapidly reaches an upper limit and does not allow further improvement despite an increase in the amount of data. On the other hand, the performance of MAP adaptation greatly improves in accordance with increasing amounts of data. But it requires considerable time in large-scale speech recognition to re-estimate model parameters. Several studies investigated ways of combining these two approaches, but the performance of the combined approaches also demonstrated great dependency on the amount of adaptation data (Goronzy and Kompe, 1999; Wang et al., 2009). The eigenvoice approach provides better performance with a relatively small amount of adaptation data. And it is capable of further improving its performance with increased amounts of data.

Another case should be considered in terms of the way adaptation data is collected. It is impractical to obtain a transcription or labels from target speakers by asking them to record sequences of specific words or sentences while collecting adaptation data. Hence, the unsupervised adaptation manner is more desirable for general voice interaction systems than the supervised manner. Among the MLLR and MAP techniques, MLLR has been characterized as better suited to unsupervised adaptation because it is robust against labeling errors (Woodland et al., 1996). Nevertheless, many studies have reported that the eigenvoice approach significantly outperforms MAP and MLLR in unsupervised conditions (Kuhn et al., 1998).

The above-mentioned general tendencies of users who interact with a device prompt the conclusion that the eigenvoice approach is more desirable for real-world voice interaction in comparison with the other approaches. For this reason, this study proposes a