Neurocomputing 129 (2014) 199–207

Contents lists available at ScienceDirect

Neurocomputing

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

Real-time frequency-based noise-robust Automatic Speech Recognition using Multi-Nets Artificial Neural Networks: A multi-views multi-learners approach

Seyed Reza Shahamiri *, Siti Salwah Binti Salim [1]

Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 21 May 2013
Received in revised form
15 August 2013
Accepted 27 September 2013
Communicated by Shiliang Sun
Available online 24 October 2013

Keywords:
Multiple-views multiple-learners
Automatic Speech Recognition
Artificial neural network
Noise robustness
Frequency-based noise

ABSTRACT

Automatic Speech Recognition (ASR) is a technology for identifying uttered word(s) represented as an acoustic signal. However, one of the important aspects of a noise-robust ASR system is its ability to recognise speech accurately in noisy conditions. This paper studies the applications of Multi-Nets Artificial Neural Networks (M-N ANNs), a realisation of multiple-views multiple-learners approach, as Multi-Networks Speech Recognisers (M-NSRs) in providing a real-time, frequency-based noise-robust ASR model. M-NSRs define speech features associated with each word as a different view and apply a standalone ANN as one of the learners to approximate that view; meanwhile, multiple-views single-learner (MVSL) ANN-based speech recognisers employ only one ANN to memorise the features of the entire vocabulary. In this research, an M-NSR was provided and evaluated using unforeseen test data that were affected by white, brown, and pink noises; more specifically, 27 experiments were conducted on noisy speech to measure the accuracy and recognition rate of the proposed model. Furthermore, the results of the M-NSR were compared in detail with an MVSL ANN-based ASR system. The M-NSR recorded an improved average recognition rate by up to 20.14% when it was given the test data infected with noise in our experiments. It is shown that the M-NSR with higher degree of generalisability can handle frequency-based noise because it has higher recognition rate than the previous model under noisy conditions.

1. Introduction

An *Automatic Speech Recognition* (*ASR*) system identifies uttered word(s) represented as an acoustic signal. An ASR system relies on a given lexicon and prior knowledge of a problem domain to recognise spoken word(s). ASR has several applications in voice-enabled control systems such as those implemented in health care, military, telephony and other domains. Nonetheless, speech recognisers are generally unable to show performance equivalent to that of human level under realistic conditions (i.e. noisy conditions). Although most of the recent speech recognisers possess high recognition rates in the lab, their performance in real-life applications under noisy environments remain unsatisfactory. This is because noise may introduce a mismatch between the data considered for ASR modelling and the actual speech data while the recogniser is being used; this mismatch of data degrades the recognition rate of the ASR system [1]. Therefore, for speech communication, accurate classification of uttered words under noisy conditions is necessary [2].

A pattern recognition system usually consists of two major components: a feature extractor and a classifier. The first component transforms an input into a representation that is trivially convertible into a class decision. Feature extractor is used to convert raw input into a form that is easily classifiable; this is a common place to incorporate classifiers such as *Artificial Neural Networks* (*ANNs*). They are mathematical models imitating natural neural systems; they have been used widely by both academia and industry as classifiers. ANNs are considered universal classifiers; to put it differently, they can in theory learn any mapping functions based on samples of inputs to the function under simulation and their responses [3].

Lately, in the studies of noise-robust ASR systems, there is a trend towards isolating noise from speech data rather than enhancing recognisers to accept and handle noisy data [4]. Nonetheless,

* Corresponding author at: BS16, Block B, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia.
Tel.: +60 3 79676300; fax: +60 3 79579249.
E-mail addresses: admin@rezanet.com (S.R. Shahamiri), salwa@um.edu.my (S.S. Binti Salim).

[1] Tel.: +60 3 79676300; fax: +60 3 79579249.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.09.040

noise reduction may not be effective when isolating noise is difficult or impossible, in addition to the fact that these methods are usually not implemented in real time. In these methods, pre-noise-processing operations need to be carried out in order to define a noise profile, detect noise, and finally isolate noise, before speech data are given to an ASR system. The drawbacks of noise isolation techniques are further discussed in Section 2.2. On the other hand, a real-time and noise-robust ASR system can recognise speech more accurately without the requirement of any noise processing. In addition, noise reduction alone may not always be a successful approach to provide clean data since it is likely that some noise will remain in the speech data after noise-robust methods are applied. Finally, noise reduction algorithms may require considerable amount of processing overhead, which makes them unsuitable for low-capacity devices. Therefore, it is important to consider approaches that can provide real-time noise-robust speech recognisers without requiring any noise processing. Such approaches should improve ASR recognition rate and generalisability so that the effects of the remaining noise are minimised.

Multi-Nets ANNs (*M-N ANNs*) are based on a novel approach proposed by Sun and Qingiu called *Multiple-Views Multiple-Learners* (*MVML*) [5,6]. The general principle of MVML is that when the function under simulation is complex due to multiple views, using multiple learners increases the classification performance. Sun conducted a survey of multi-view learning theories and discussed its usefulness; he concluded that MVML is an effective approach with widespread applicability [7].

In the context of M-N ANNs, each ANN learns one of the views and together they form an ensemble of learners. They have been proven capable of recognising complex patterns better than *Multiple-Views Single-Learners* (*MVSL*) ANN-based classifiers because they distribute the complexity of the function among several parallel and independent neural networks, making the overall classification easier [8]; under similar conditions, an MVSL ANN-based system is unable to learn the function with adequate accuracy.

In this paper, we investigated whether M-N ANNs can be used to increase the recognition rate and accuracy of an ASR system in noisy conditions. A speaker-independent *Multi-Network Speech Recogniser* (*M-NSR*) is provided using M-N ANNs, in which each neural network represents one of the words in the vocabulary. The model proposed here assumes that speech data are corrupted with noise, regardless of application of a noise robustness method. The model adopts a real-time approach, which means that it does not need to perform any specific noise processing or digital signal processing to isolate or compensate for the noise; instead, we concentrated on providing a more accurate ASR model that identifies noisy acoustic signals with a better recognition rate. We also experimented with the helpfulness of increasing the ASR model's generalisability to provide a frequency-based noise-robust ASR model. For evaluation purpose, the trained M-NSR was used to identify unforeseen test data that were affected by frequency-based white, brown and pink noises, each having a different Signal-to-Noise Ratio (SNR).

Since this paper studies the advantages of MVML ANN-based ASR system over MVSL, an MVSL ANN-based ASR system was provided as the reference model. It was trained and tested with the same data as well as same methodology. Finally, the results obtained from the proposed M-NSR model and the reference model were meticulously compared.

2. Related work

This section is divided into two parts. The first part briefly surveys the applications of MVSL ASR systems based on ANNs. The second part highlights some of the state-of-the-art approaches in building noise-robust speech recognisers.

2.1. ANN applications in ASR Systems

ANNs have been used for speech recognition since the late 1980s. An MVSL ANN-based ASR system is any speech recogniser that uses a single ANN to map speech features into words or phonemes, such as the one used by [9]. Among the early successful attempts, Lang et al. [10] proposed a *Time-Delay Neural Network* (*TDNN*) architecture for phoneme recognition. TDNNs are used for real-time, continuous data adaption, in which inputs are delayed and then fed to the ANN from the leftmost set of neurons of the input layer. Neurons of the output layer and hidden layer(s) are replicated over time. The authors used this method to design a speaker-independent recogniser and obtained 92.2% accuracy in identifying “B”, “D”, “E”, and “V”. Similarly, ANNs were used in the following applications: Japanese phoneme classifier [11], Malay vowel recognition [12], Hindi consonants recognition [13,14], and fuzzy vector quantiser for Arabic speech recognition [15].

Nejadgholi and Seyyedsalehi [16] applied ANNs to address the speaker variation problem in ASR systems. Speaker variation may cause input signals of an ASR system to change, which in turn reduces the system accuracy. In order to mitigate the problem, they applied four neural networks. The first network was a TDNN used to identify a speaker as the “referenced speaker”. The second network was trained using the referenced speaker data. The third network provided “adapted features”, which means it performed an inverse mapping of the phoneme classification and speaker data into the input signal that may adapt the input representation to the identified speaker. The final network employed the adapted features to make phoneme recognition. The authors claimed that accuracy of phoneme recognition is increased by up to 3%. However, performance is highly dependent on the success of the system in identifying the referenced speaker because if the speaker is misidentified, the accuracy may decrease drastically.

2.2. State-of-the-art speech noise-robustness methods

Noise robustness is one of the major concerns in recent ASR studies. Some previous studies of robust ASR systems recommended using clean speech data for ASR training. The disadvantage of this method is that, whenever a recogniser is used in a noisy environment, noisy speech data and clean data may produce different features that can result in a huge impact by lowering the recognition rate of an ASR system. In order to deal with this problem, other studies (such as [17]) suggested that noisy data should be considered for training instead of clean data. However, since different noise sources produce different types of noises, this method is not effective enough because it is not possible to consider all types of noises.

Therefore, other approaches that specifically process noise have been considered. There are two major categories of speech noise robustness. The first one considers noise caused by missing or incomplete speech data, which is called *noise compensation* [18,19]; this means that some components of features that represent a speech signal are not available. Cooke et al. proposed an approach to resolve this problem by integrating two methods, namely *data imputation* and *marginalisation* [18]. Data imputation, which replaces missing data with clean estimates, is applied in order to substitute the missing components; marginalisation is used to identify components that are still missing after data imputation has been applied. Imputation for a large vocabulary was studied extensively by [20]. Examples of researchers who studied noise compensation are as follows: Mohammadi and Almasganj [19], who considered *multivariate Laplacian distribution*