International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 6, December 2017, pp. 3358 – 3368
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i6.pp3358-3368

Recent advances in LVCSR : A benchmark comparison of performances

Rahhal Errattahi and Asmaa El Hannani
Laboratory of Information Technology, National School of Applied Sciences, University of Chouaib Doukkali, EL Jadida, Morocco

Article Info

Article history:
Received: Nov 26, 2016
Revised: Jun 28, 2017
Accepted: Jul 10, 2017

Keyword:
Large Vocabulary Continuous Speech Recognition
Automatic Speech Recognition
Word Error Rates
Deep Neural Networks
Hidden Markov models
Gaussian Mixture Models

ABSTRACT

Large Vocabulary Continuous Speech Recognition (LVCSR), which is characterized by a high variability of the speech, is the most challenging task in automatic speech recognition (ASR). Believing that the evaluation of ASR systems on relevant and common speech corpora is one of the key factors that help accelerating research, we present, in this paper, a benchmark comparison of the performances of the current state-of-the-art LVCSR systems over different speech recognition tasks. Furthermore, we put objectively into evidence the best performing technologies and the best accuracy achieved so far in each task. The benchmarks have shown that the Deep Neural Networks and Convolutional Neural Networks have proven their efficiency on several LVCSR tasks by outperforming the traditional Hidden Markov Models and Guaussian Mixture Models. They have also shown that despite the satisfying performances in some LVCSR tasks, the problem of large-vocabulary speech recognition is far from being solved in some others, where more research efforts are still needed.

Corresponding Author:
Rahhal Errattahi
Laboratory of Information Technology, National School of Applied Sciences, University of Chouaib Doukkali, EL Jadida, Morocco
errattahi.r@ucd.ac.ma

1. INTRODUCTION

Speech is a natural and fundamental communication vehicle which can be considered as one of the most appropriate media for human-machine interactions. The aim of Automatic Speech Recognition (ASR) systems is to convert a speech signal into a sequence of words either for text-based communication purposes or for device controlling. ASR is usually used when the keyboard becomes inconvenient such, for example, when our hands are busy or with limited mobility, when we are using the phone, we are in the dark, or we are moving around etc. ASR finds application in many different areas: dictation, meeting and lectures transcription, speech translation, voice-search, phone based services and others. Those systems are, in general, extremely dependent on the data used for training the models, configuration of front-ends etc. Hence a large part of system development usually involves investigations of appropriate configurations for a new domain, new training data, and new language.

There are several tasks of speech recognition and the difference between these tasks rests mainly on: (i) the speech type (isolated or continuous speech), (ii) the speaker mode (speaker dependent or independent), (iii) the vocabulary size (small, medium or large) and (iv) the speaking style (read or spontaneous speech). Even though ASR has matured to the point of commercial applications, the Speaker Independent Large Vocabulary Continuous Speech Recognition tasks (commonly designed as LVCSR) pose a particular challenge to ASR technology developers. Three of the major problems that arise when LVCSR systems are being developed are: First speaker independent systems require a large amount of training data in order to cover speakers variability. Second, continuous speech recognition is very complex because of the difficulties to locate word boundaries and the high degree of pronunciation variation due to dialects, coarticulation and noise, unlike isolated word

speech recognition where the system operates on single words at a time [1, 2, 3]. Finally, with large vocabulary, it becomes increasingly harder to find sufficient data to train the acoustic models and even the language models. Thus, subwords models are usually used instead of words models which affect negatively the performance of the recognition. Moreover, LVCSR tasks themselves vary in difficulty; for example read speech task (human-to-machine speech, e.g. dictation) is much easier than spontaneous speech task (human-to-human speech, e.g. telephone conversation).

To deal with all these problems, there has been a plethora of algorithms and technologies proposed by the scientific communities for all steps of LVCSR over the last decade: pre-processing, feature extraction, acoustic modeling, language modeling, decoding and result post-processing. Many papers were dedicated to presenting an overview of the advances in LVCSR: [4, 5, 6, 7, 8]. However, the scope of this paper focuses primarily on the systems architecture, the techniques used and the key issues. There is to date no work which has attempted to report and analyze the current advances in term of performances across all the major tasks of LVCSR. Our ambition has been to fill this gap. We did a benchmark comparison of the performances of the current state-of-the-art LVCSR systems and have covered different tasks: read continues speech recognition, mobile voice search, conversational telephonic speech recognition, broadcast news speech recognition, video speech and distant conversational speech recognition. We tried to put objectively into evidence the best performing technologies and the best accuracy achieved so far in each task. Note that in this paper we only address the English language and that we have constrained the review to systems that have been evaluated on widely used speech corpora. This choice was forced by the non-availability of enough publication of some other corpora and also because evaluating systems on relevant and common speech corpora is a key factor for measuring the progress and discovering the remaining difficulties and especially when comparing systems produced by different labs.

Table 1. Overview of the selected LVCSR corpora

Corpus	Year	Type of data	Size	Audio source
Wall Street Journal I, [9]	1991	Read excerpts from the Wall Street Journal	80 hours, 123 speakers	Close talking microphone
Switchboard, [10]	1993	Phone conversations between strangers on an assigned topic	250 hours, 543 speakers, 2400 conversations	Variable telephone handsets
CallHome, [11]	1997	Phone conversations between family members or close friends	120 conversations, Up to 30 min each	Variable telephone handsets
Broadcast News, [12, 13]	1996/1997	Television and radio broadcast	LDC97S44:104 hours LDC98S71:97 hours	Head mounted microphone
AMI, [14]	2007	Scenario and NonScenario Meetings from various groups	100 hours	Close-talking and far-field microphones
Bing Mobile Data, [15]	2010	Mobile voice queries	21400 hours	Variable mobile phones
Google Voice Search Data	-	Mobile Voice Search and Android Voice Input	5780 hours	-
Youtube Video	-	Video from youtube	1400 hours	-

2. COMPARING STATE-OF-THE-ART LVCSR SYSTEMS PERFORMANCES

The industry and research community can benefit greatly when different systems are evaluated on a common ground and particularly on the same speech corpora. In this perspective, we report the recent progress in the area of LVCSR on a selection of most popular English speech corpora with vocabularies ranging from 5K to more than 65K words and content ranging from read speech to spontaneous conversations. Only such with recent publications were considered. An overview on properties of the chosen sets is given in Table 1. In the following subsections, we will shortly introduce each task and the datasets used and report the performances of systems produced by different labs.

2.1. Read continuous speech recognition task

Early speech recognition systems were often designed for read speech transcription tasks (dictation). As implied by the name, the data used in this domain consists of read sentences and in general in a speaker-independent mode. Its popularity arose because the lexical and syntactic content of the data can be controlled and it is significantly less expensive to collect than spontaneous speech. The primary applications in this domain include the dictation of notes and transcription of important information by some professionals (e.g. medical, military and law) and by persons with learning disabilities (e.g. dyslexia and dysgraphia), limited motor skills or vision impairment.

The Wall Street Journal corpus I (also known as CSR-I or WSJ0) [9] is known as a reference corpus in the field, which is an American English read speech with texts taken from the Wall Street Journal news, the speech was recorded using a machine-readable under clean conditions. The systems presented here were evaluated on the November 1992 ARPA CSR (Nov-92) benchmark test set, a 5K-word closed- vocabulary subset derived from the WSJ0 corpus which consists of 330 utterances from 8 speakers. Hidden Markov models (HMMs) and Guaussian Mixture Models (GMMs) have been used extensively since the beginning of the research in the area of speech recognition. More than 40 years later, they still predominate and they are usually used as baseline when it comes to compare systems with different acoustic models. Beside this, several techniques were developed around the HMMs/GMMs in order to improve the performance of the ASR systems.

Table 2. Word Error Rates (WER) in % on the Nov-92 subset of the WSJ0 corpus using bigram and trigram language models

Acoustic model / Features	Bigram	Trigram
MLP-HMM, [16]	8.5	6.5
RC-HMM, [17]	6.2	3.9
GMM-HMM (ML), [17]	6.0	3.8
GMM-HMM (MMI+VTLN), [16]	-	3.0
DNN-HMM (STC features), [18]	5.2	-

Table 2 shows a recapitulation of the key performances of some state-of-the-art systems in the field. The fisrt two systems in the list are based on a GMM-HMM acoustic models: the first one was trained using the maximum-likelihood (ML) criteria [17], while the second one using the maximum mutual information (MMI) criteria with the vocal tract length normalization (VTLN) [16].Triefenbach et al. [17], proposed also a Reservoir Computing (RC) HMM hybrid system for phoneme recognition using a bigram phonotactic utterance model. The RC-HMM performs significantly better than the MLP-HMM hybrids proposed by Gemello et al. [19]. However, it is still outperformed by the GMM system with VTLN. Another study [18] demonstrated the effectiveness of the Deep Neural Networks (DNNs) in speech recognition. The best result of this study belongs to a DNN system with 5 hidden layers, where each hidden layer has 2048 nodes. In terms of complexity, both the DNN-HMM and the RC-HMM incorporates massive parameters in the training stage. On the other side, GMM-HMM is much more efficient as it reaches good performances with small number of parameters.

The results, obtained with the bigram language model, shows that the DNN-HMM acoustic model presents the best performance on the WSJ0 task; this performance could be even enhanced using a trigram language model. Generally, using trigram language model was crucial and clearly superior to using bigram models over the Nov-92 test set in various studies.

2.2. Voice search speech recognition task

Voice search is the technology allowing users to use their voice to access information. The advent of smart phones and other small, Web-enabled mobile devices in recent years has spurred more interest in voice search, especially in some usage scenarios when our hands are busy or with limited mobility, when we are using the phone, we are in the dark, or we are moving around etc. There is a plethora of mobile applications which allows users to give speech commands to a mobile either for a search purpose (e.g web search, maps, directions, travel resources such as airlines, hotels etc) or for question answering assistance purpose. Mobile voice search speech recognition is consedered as one of the challenging tasks is the field of speech recognition due to many factors: the utterances tend to be very short, yet unconstrained and open-domain. Hence, the

vocabularies are unlimited with unpredictable input, and high degree of acoustic variability caused by noise, side-speech, accents, sloppy pronunciation, hesitation, repetition, interruptions, and mobile phone differences.

In this section we will report results on two voice search application that have been built in the recent few years: the Google Voice Input and the Bing mobile voice search.

2.2.1. Google Voice Search

Google Voice Search transcribes speech input used for user interaction from mobile devices as voice search queries, short messages and emails. The Google Voice Search system was trained using approximately 5780 hours of data from mobile Voice Search and Android Voice Input.

Table 3. WER in % on a test set from the Google live Voice input dataset

Acoustic model	WER
GMM-HMM, [20]	16.0
DNN(DBN)-HMM, [20]	12.3
+ MMI discriminative training, [20]	12.2
DNN + GMM (combination), [20]	11.8

The baseline GMM-HMM [20] system created by Google's group consisted of triphone HMM with decision tree clustered states, and used PLP features that were transformed by linear discriminative-Analysis (LDA). The GMM-HMM model was trained discriminatively using Boosted-MMI criterion. All these parameters generate a context-dependent model with a total of 7969 states. The same data model was used to train a deep belief networks (DBN) based DNN acoustic model to predict the 7969 HMM states. The used DBN-DNN in [20] was composed of four hidden layers with 2560 nodes per each hidden layer, a final layer with 7969 states, and an input layer of 11 contiguous frames of 40 log filter-bank were modelled with a DBN.

As shown in Table 3 the DNN(DBN)-HMM system achieved a 23% relative reduction over the baseline. Further improvement could result from combining both systems using the segmental conditional random field (SCARF) framework, this combination gave a word error rate of 11.8%.

2.2.2. Bing mobile voice search

Bing mobile voice search, known as Live Search for mobile (LS4M) [15], is a mobile application that allows users to do web-based search (e.g. map, directions, traffic, and movies) from their mobile phones. LS4M was developed by the Microsoft company and was trained on a data set around 24 hours with a high degree of acoustic variability.

Table 4. WER in % on the Bing Voice Search set

Acoustic model	WER
DNN-HMM (No PT), [21]	37.1
DNN-HMM (with PT), [21]	35.4
CNN-HMM (No PT), [21]	34.2
CNN-HMM (with PT), [21]	33.4
RNNLM, [22]	23.2

Abdel-Hamid et al. [21] at Microsoft Research, investigate the performance of both DNNs and CNNs on a LVCSR task and the effects of RBM-based pretraining on their recognition performance. Both models were trained using a subset of 18 hours from the Bing Voice Search task in order to predict the triphone HMM state labels. The DNNs architecture consisted of three hidden layers while the CNN had one pair of convolution and pooling plies in addition to two hidden fully connected layers. Results, in Table 4 show that the CNN outperform the DNN on the Bing Voice Search, providing about 8% relative error reduction without pretraining and relative word error rate reduction of 6% while using pretraining. According to Abdel-Hamid et al. [21], pretraining is more effective for the CNN than for the DNN. Another study [22], suggests applying Recurrent Neural Network Language Models (RNNLMs) directly in the first pass of speech recognition decoding, which outperform both simple n-gram based models (DNN and CNN) on the Bing Voice Search task with a word

error rate of 23.2 %. However, the computational expense of RNNLMs is very high, and to reduce the cost of using a RNNLMs, authors propose cache based RNN inference, which drops the runtime from 100xRT (no caching is done) to just under 1.2xRT.

Though the experimental setup was not described in sufficient detail, in both papers, we can only assume that the 10% absolute improvement of WER in RNNLM vs CNN systems could be due to differences in the amount of training data or differences of the Bing Voice Search subset used to evaluate the systems.

2.3. Conversational telephone speech recognition task

Owing to the revolution in telecommunication domain, peoples all over the world spent millions of hours in communication via their phones. For many reasons, as security for example, the transcription of spontaneous casual speech and particularly conversational telephone speech becomes indispensable. Whereas transcribing this type of speech is very challenging, due to many factors, including poor articulation, increased coarticulation, highly variable speaking rate, and various types of disfluency such as hesitations, false starts, and corrections.

We report pertinent results on a highly challenging test set, the NIST 2000 Hub5 (Hub5'00). The Hub5 is composed of two subsets, an "easy" split which containe 20 conversation from Switchboard corpus [10, 23], and a "hard" split containing 20 conversation from CallHome corpus [11], often reporting results on the easier portion alone. Switchboard is a corpus of American English spontaneous conversational telephone speech, it is composed of about 2,400 two-sided telephone conversations between 543 speakers (302 male, 241 female) from all areas of the United States. While the CallHome corpus consists of 120 unscripted telephone conversations between native speakers of English mostly between family members or close friends overseas. The CallHome data is harder to recognize compared to Switchboard, partly due to a greater presence of foreign-accented speech.

In the last two years, several labs have conducted benchmarking experiements using the Switchboard corpus [24, 25, 26, 27, 28, 29]. In Table 5, we summarize the most performing sytems on the Hub5'00 dataset splits. All systems have been trained on the 300 hour Switchboard dataset except the Deep speech system from [26] which has been trained on both Switchboard and Fisher dataset. The Fisher corpus [30] offers 2000 hours of conversational telephone speech collected in a similar manner as Switchboard.

Table 5. WER in % on the Switchboard subsets "SWB" of the Hub5'00 dataset

Acoustic Models	SWB
GMM/HMM fBMMI, [31]	14.5
DNN-HMM-sMBR fMLLR, [24]	12.6
RNN (Deep speech), [26]	12.6
DNN fMLLR, [31]	12.2
CNN log-mel, [31]	11.8
CNN+DNN log-mel+fMLLR+I-vector, [31]	10.7
MLP/CNN+I-Vector, [28]	10.4

The GMM system in [31], was trained using speaker-adaptation with VTLN and feature space Maximum Likelihood Linear Regression (fMLLR), followed by feature and model-space discriminative training with the the Boosted Maximum Mutual Information (BMMI) criterion. The DNN-HMM sMBR system from [24] was trained on LDA+STC+FMLLR features on the full 300 hour training set, and was composed of 7 layers, where each hidden layer has 2048 neurons; an an output layer of 8859 units. Hannun et al. [26] propose a RNN-based system called Deep Speech that uses deep learning systems to learn from large datasets (more than 7380 hours). In [26], authors used a multi-GPU computation for training the RNN model, and a combination of collected and synthesized data, which make the system able to learn robustness to realistic noise and speaker variation. The DNN system, in [31], was trained using fMLLR features and was composed of 6 hidden layers each containing 2048 sigmoidal neurons, and a softmax layer with 8192 output units. For the CNN [31], it was trained using log-mel features, with an architecture consisted of two convolutional layers each containing 512 hidden units, and five fully connected layers each containing 2048 hidden units, and a softmax layer with 8192 output units. Results are summarised in Table 5, show that the CNNs clearly outperform the other systems giving a 20% word error rate relative improvement over the GMM/HMM system, and 3% word error rate relative

improvement over the hybrid DNN. Another form of system combination have been proposed in [28]: a jointly trained MLP/CNN model with I-Vectors, where the outputs of the first MLP hidden layer get combined with the outputs of the second CNN layer. This system has given the best result in this task so far (10.4% WER on the SWB split of Hub5'00).

2.4. Broadcast news speech recognition task

Broadcast News Automatic Speech Recognition (ASR BN) consist of recognising speech from news-oriented content from either television or radio, including news, multi-speaker roundtable discussions, debates, and even open-air interviews outside of the studio. English Broadcast News Speech Corpus [12, 13] is one of the most common datasets in this domain. It is a collection of radio and television news broadcasts (from ABC, CNN and CSPAN television networks and NPR and PRI radio networks) with corresponding transcripts.

The acoustic models of the systems reviewed in this section was trained on 50 hours of data from the 1996 (LDC97S44, 104 hours) and 1997 (LDC98S71, 97 hours) English Broadcast News Speech Corpora. State-of-the-art systems performances are reported on both the EARS Dev-04f (3 hours from 6 shows) and RT-04 (6 hours from 12 shows) sets.

Table 6. WER in % on the Dev-04f and RT-04 English Broadcast News sets

Acoustic model	Dev-04f	RT-04
Baseline GMM/HMM, [27]	18.8	18.1
SAT MLP fBMMI, [32]	21.9	23.6
SAT DBN fBMMI, [32]	17.0	17.7
SAT GMM fMPE+MPE, [33]	16.5	14.8
SAT DNN cross-entropy, [33]	16.7	14.6
SAT DNN HF sMBR, [33]	15.1	13.4
Hyprid DNN, [27]	16.3	15.8
DNN-based Features, [27]	16.7	16.0
Hybrid CNN, [27]	15.8	15.0
CNN-based Features, [27]	15.2	15.0

In [27], the baseline GMM-HMM system was trained using speaker-based mean with VTLN and an LDA transform to project the 13-dimentional MFCC to 40 dimensions. Next a FMLLR followed by a BBMI transform were applied to obtain a GMM system with 2220 quinphone states and 30k diagonal covariance Gaussians. Sainath et al. [32] compared the performance of Deep Belief Networks (DBNs) to a simple Multi-Layer Perceptrons MLPs, where both the DBN and MLP were trained with the same architecture (6 layers x 1,024 units) using speaker-adapted and fBMMI features with 2220 output targets. The DBN contains two types of Restricted Boltzmann Machines (RBMs) that was used to pre-train weights of the Artificial Neural Networks (ANNs). For the first layer of DBN authors used a Gaussian-Bernoulli RBM trained for 50 epochs. The first layer consist of 9 frames as inputs and 1,024 output features. For all subsequent layers, Bernoulli-Bernoulli RBMs are trained for 25 epochs and contain 1,024 hidden units. Further study of Sainath et al. [33] suggests different refinements to improve DNN training speed. The DNNs systems use a fMLLR with 5,999 quinphone states, and composed of six hidden layers each containing 1,024 sigmoidal units. In this study, authors succeeded in reducing the number of parameters from 10.7 M to 5.5 M a 49% reduction, due to low-rank factorization. Recent study of Sainath et al. [27] explored the performance of the CNNs compared to DNN. The Hybrid DNN has an architecture of 5 layers with 1024 hidden units per each layer and a softmax output layer with 2220 target units. The DNN-based system has the same architecture, but with only 512 output targets. While for booth Hybrid CNN and CNN-based feature systems are trained using VTLN-warped mel FB, delta and double-delta feature.

Table 6 shows that RBM pre-training of the DBN improves the WER over the MLP for all feature sets. Following sMBR training, the DNN is the best model. It is 20% better than the baseline GMM on *Dev-04f* and 36% better on *RT-04*.Furthermore, the CNN-based features present competitive performance with 19% relative improvement over the baseline GMM-HMM. The performance of CNN-based features could achieve only 13.1% WER on the dev04 and 12.0% WER on TR04 [27], when a larger scale tasks is used for training (400 hours of English Broadcast News).

2.5. Video speech recognition task

The advent of the web, low cost digital cameras, and Smatphones has significantly broadened the quantity as well as the reach of videos. The key challenge for many web video producers is making it easy for others, hearing impaired and non-native speakers, to find and enjoy their content. One way to do that is to use hand-transcription, even so this solution can be time-consuming and expensive, and could not cope up with the huge content being uploaded every minute to the internet. On the other hand automatic video transcription represents an alternative solution to improve accessibility of the video contents. In this section we chose to present results studies done by the Google researchers on the Youtube video data. The goal of this task is to transcribe Youtube data, unlike the previous tasks YouTube data is extremely challenging for current ASR technology [34].

Jaitly et al. [20] used 1400 hours of YouTube data to train the Context-Dependent ANN/HMM with speaker adapted features and 17552 triphone target states. The baseline system used 9-frames of MFCCs as inputs that were transformed using LDA. The acoustic models were further improved with BMMI. During decoding, fMLLR and MLLR transforms were applied. For the DBN-HMMs the acoustic data used in the training stage were the fMLLR transformed features. For a complexity reason and to make the training faster, the ANN/HMM has an architecture of only 4 hidden layers with 2000 units in the output layer and 1000 units in the layers above.

In order to generate additional semi-supervised training data, Liao et al. [34] have proposed to use the owner-uploaded video transcripts and a DNNs acoustic models. The proposed DNNs are fully-connected, feed forward neural network, with sigmoid non-linearities and a softmax output layer and was trained using minibatch stochastic gradient descent and back-propagation techniques.

Reported results are summarized in Table 7. It should be noted that the training set used in the experiments is not exactly the same, but both experiments was conducted on a comparable amount of data; namely 1400 hours in [20] vs 1781 hours in [34]; however, al results are reported on the same test set- YtiDev11 (6.6 hours, 2.4M frames). The reported baseline 7x1024 system with 7k output states from [34] outperform all of the these previously reported results in [20]. While merging the wide hidden layer architecture of 2048 nodes with a low-rank approximation with high number of CD states in the output layer yielded the best result on the YtiDev test set of 40.9% WER.

Table 7. WER in % one the YtiDev11 YouTube set

Acoustic model	WER
MFCC GMM, 18k state, 450kcomps, [20]	52.3
DBN-HMM pretrained with sparsity, [20]	47.6
+ MMI, [20]	47.1
+ system combination with SCARF, [20]	46.2
Fbank DNN 7x1024, 7k state, [34]	44.0
Fbank DNN 6x2048, 7k state, [34]	42.7
Fbank DNN 7x1024, low-rank 256, 45k state, [34]	42.5
Fbank DNN 7x2048, low-rank 256, 45k state, [34]	40.9

2.6. Distant conversational speech recognition task

Distant conversational speech recognition (DCS) is captured using multiple distant microphones, typically configured in a calibrated arrayis, and is very challenging since the speech signals to be recognized are degraded by the presence of overlapping talkers, background noise, and reverberation. Classroom Lectures, Parliamentary Meetings, and Scientific Meeting are the main applications of Distant conversational speech recognition. In this section we chose to report results over the AMI Meeting corpus, prompted by the large use of this corpus in several recent studies.

The AMI corpus [14] contains around 100 hours of meeting recordings from three European sites (UK, Netherlands, Switzerland). Each meeting usually has four participants and the meetings are in English, many of the meeting participants are non-native English speakers. The AMI corpus was divided into train, development, and test sets. Where about 78 hours of meeting recorded speech were used as training set, and about 9 hours each were used as development and test sets. Evaluations in the meeting domain are usually

conducted in three conditions: Single Distant Microphone (SDM), Multiple Distant Microphones (MDM) and Individual Headset Microphones (IHM).

Table 8. WER in % on the AMI set for various microphone configurations; SDM, MDM and IHM are respectively Single Distant Microphone, Multiple Distant Microphones and Individual Headset Microphones

	Dev			Test		
Acoustic model	SDM	MDM	IHM	SDM	MDM	IHM
GMM LDA+STC, [35]	63.2	54.8	29.4	67.66	59.4	31.6
DNN LDA+STC, [35]	55.4	51.4	26.7	59.8	56.0	28.4
DNN Fbank, [35]	55.8	51.1	28.3	60.8	55.6	31.5
CNN, [36]	52.5	46.3	25.6	-	-	-

Swietojanski et al. [35] applied DNN-HMM for meeting speech recognition task. Authors compared there results to the conventional system based on GMMs. The baseline GMM-HMM system have a total of 80000 Gaussians, and were discriminatively trained using BMMI with linear discriminative analysis (LDA) and decorrelated using a semi-tied covariance (STC) transform. While the DNNs were configured to have 6 hidden layers, with 2048 units in each hidden layer and was trained using RBMs and using either LDA+SAT features or the FBANK features. Further study of Swietojanski et al. [36], suggests using CNNs for large vocabulary distant speech recognition trained using the three type of microphones: SDM and MDM or simply IHM. The CNNs were trained using the Fbank features, and composed of a single CNN layer followed by fives fully-connected layers.

Table 8 shows that replacing the GMM with a DNN improves recognition accuracy for speech recorded with distant microphones. While in [36], they found that CNNs improve the WER 6.5% relative compared to conventional deep neural network (DNN) models and 15.7% over a discriminatively trained GMM baseline.

3. DISCUSSION

From the first view at the reported results over the different tasks using various acoustic models, we can illustrate that the traditional GMM based HMM models has been outperformed by several other models. Despite their ability to model the probability distributions over vectors of input features that are associated with each state of an HMM, GMMs have a serious shortcoming. As Hinton et al. [37] stated, "Despite all their advantages, GMMs have a serious shortcoming; they are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space". Therefore other classifiers, which can capture better properties of acoustic features could presente better accuracy than GMMs for acoustic modeling of speech.

Machine learning algorithms are more efficient than the traditional GMM for acoustic modeling of speech. In particular the neural networks (NNs) based technology present challenging performance over a variety of speech recognition benchmarks. Their successes, in acoustic modeling of speech, come from there capability to classify data even with small number of parameters and the potential to learn much better models of data that lie on or near a nonlinear manifold. DNNs, which are a feed-forward artificial neural network that has more than one layer of hidden units between its inputs and its outputs, are the new generation of the NNs, they come to solve problems of training time and overfitting by adding an initial stage of generative pre-training using RBMs.

Performances of LVCSR systems varied from domain to other, as summarized in Figure 1. In some domains, like read continuous speech where generally the speech was recorded under clean conditions, results are satisfying with an error rate under 5%. While in other domains that contain more speech variations, as video speech or distant conversational speech (meeting), results are not acceptable presenting an error rate near 50%. This huge difference in performances was caused by the nature of speech, the more natural and spontaneous the speech is, the more the error rate increase. Difficulties encountered in modeling spontaneous speech stem from many factors: foreign accents, extraneous words, out-of-vocabulary words, ungrammatical sentences, disfluency, partial words, repairs, hesitations, repetitions, style shifting It must be said that in this paper we have constrained the benchmarks to the performances in terms of word error rates, because the majority of researchers use it as common measure to report performance of their systems. However, other important aspects of ASR systems should also be taken into account in the future, such as the efficiency and the usability. Most of the systems presented in the literature require either lots of training data (thousands of

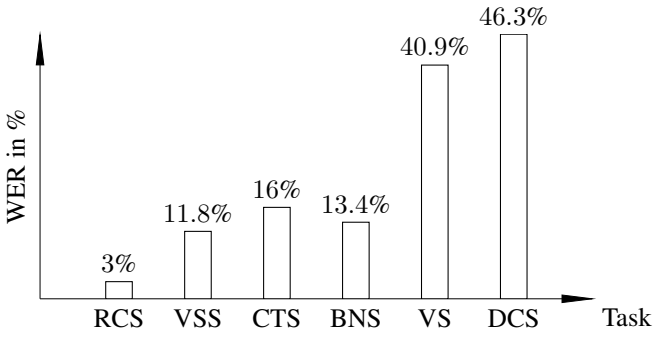

Figure 1. Best state-of-the-art performances over the six tasks. RCS, VSS, CTS, BNS, VS, DCS are respectively read continuous speech, voice search speech, conversational telephone speech, broadcast news speech, video speech and distant conversational speech.

hours of speech and billions of words of text) or large computational expense which is ineffective. Therefore, we believe there is a need of corpora and evaluations that include more objective criteria, oriented towards usability, in order to develop a more user-centered ASR application. It should be noticed that the ASR system must ensure reactiveness; looking at the real time factor of the used algorithms, and robustness; in front of accents and impaired speech should also be considered.

4. CONCLUSION

In this paper we have summarized the recent developments of LVCSR research and presented a benchmark comparison of the performances of ASR systems on different LVCSR tasks: read continues speech recognition, mobile voice search, conversational telephonic speech recognition, broadcast news speech recognition, video speech and distant conversational speech recognition. Most of the presented results show that replacing GMMs with other machine learning algorithms gives competitive results. Particularly, the DNN gives fascinating performances over a variety of speech recognition benchmarks. The biggest disadvantage of DNNs is there complexity; it is hard to train a large model on massive datasets. Although we suggest that any improvement for a clean speech corpus such as WSJ is promising. On the other hand more potential researches are needed in several domains that are characterized by noisy and spontaneous speech such as video and distant conversational speech.

REFERENCES

[1] T. Adam, et al., "Wavelet cesptral coefficients for isolated speech recognition," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 5, pp. 2731–2738, 2013.
[2] N. R. Emillia, et al., "Isolated word recognition using ergodic hidden markov models and genetic algorithm," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 10, no. 1, pp. 129–136, 2012.
[3] F. Jalili and M. J. Barani, "Speech recognition using combined fuzzy and ant colony algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 5, pp. 2205–2210, 2016.
[4] S. Young, "A review of large-vocabulary continuous-speech," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45–57, Sept 1996.
[5] G. Zweig and M. Picheny, "Advances in large vocabulary continuous speech recognition," *Advances in Computers*, vol. 60, pp. 249–291, 2004.
[6] G. Saon and J.-T. Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.
[7] J. Baker, et al., "Developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, 2009.
[8] J. Baker, et al., "Updated minds report on speech recognition and understanding, part 2," vol. 26, no. 4, 2009, pp. 78–85.
[9] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *DARPA Speech and Language Workshop*. Morgan Kaufmann Publishers, 1992.

[10] J. Godfrey and E. Holliman, "Switchboard-1 release 2 ldc97s62," 1993.
[11] A. Canavan, et al., "Callhome american english speech ldc97s42," *Linguistic Data Consortium, Philadelphia:*, 1997.
[12] J. Fiscus, et al., "1997 english broadcast news speech (hub4) ldc98s71," *Linguistic Data Consortium, Philadelphia*, 1997.
[13] e. a. Graff, David, "1996 english broadcast news speech (hub4)ldc97s44," *Linguistic Data Consortium, Philadelphia*, 1996.
[14] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.
[15] A. Acero, et al., "Live search for mobile:web services by voice on the cellphone," in *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2008, pp. 5256–5259.
[16] G. Heigold, et al., "Discriminative hmms, log-linear models, and CRFS: what is the difference?" in *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2010, pp. 5546–5549.
[17] F. Triefenbach, K. Demuynck, and J.-P. Martens, "Large vocabulary continuous speech recognition with reservoir-based acoustic models," *IEEE Signal Processing Letters*, vol. 21, no. 3, pp. 311–315, March 2014.
[18] S. Siniscalchi, T. Svendsen, and C.-H. Lee, "A bottom-up modular search approach to large vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 786–797, April 2013.
[19] R. Gemello, et al., "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.
[20] N. Jaitly, et al., "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proceedings of Interspeech*, 2012.
[21] O. Abdel-Hamid, et al., "Convolutional neural networks for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.
[22] Z. Huang, et al., "Cache based recurrent neural network language model inference for first pass speech recognition," in *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014, pp. 6354–6358.
[23] J. J. Godfrey, et al., "Switchboard: Telephone speech corpus for research and development," in *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 1992, pp. 517–520.
[24] L. B. K. Vesely, et al., "Sequence-discriminative training of deep neural networks," in *In Interspeech*, 2013.
[25] F. Seide, et al., "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, December 2011.
[26] A. Y. Hannun, et al., "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.
[27] T. N. Sainath, et al., "Deep convolutional neural networks for lvcsr," in *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8614–8618.
[28] H. Soltau, et al., "Joint training of convolutional and non-convolutional neural networks," *In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.
[29] A. L. Maas, et al., "Increasing deep neural network acoustic model size for large vocabulary continuous speech recognition," *arXiv preprint arXiv:1406.7806*, 2014.
[30] C. Cieri, et al., "The fisher corpus: a resource for the next generations of speech-to-text," in *LREC*, vol. 4, 2004, pp. 69–71.
[31] T. N. Sainath, et al., "Deep convolutional neural networks for large-scale speech tasks," *Elsevier, Special Issue in Deep Learning*, 2014.
[32] T. Sainath, et al., "Making deep belief networks effective for large vocabulary continuous speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2011, pp. 30–35.

[33] T. Sainath, et al., "Optimization techniques to improve training speed of deep neural networks for large speech tasks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2267–2276, Nov 2013.
[34] H. Liao, et al., "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2013, pp. 368–373.
[35] P. Swietojanski, et al., "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2013, pp. 285–290.
[36] P. Swietojanski, et al., "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, Sept 2014.
[37] G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

BIOGRAPHY OF AUTHORS

Rahhal Errattahi is a Ph.D student at Laboratory of Information Technology, National School of Applied Sciences, University of Chouaib Doukkali, EL Jadida (Morocco). He obtained Master Degree in computer science specialty of Business Intelligence from the University of Sultan Moulay Slimane (Morocco) in 2014. His researches are in fields of speech recognition, data mining, data analysis and natural language processing. He prepare a dissertation on the automatic detection and correction of speech recognition errors using data mining techniques.

Dr Asmaa El Hannani is an Assistant Professor in Computer Science at the National School of Applied Sciences, University of Chouaib Doukkali, EL Jadida (Morocco). She received a Diploma degree (MSc) in computer science from the University of Fribourg (Switzerland), in 2003. In 2007, she obtained a joint PhD degree in computer science from the University of Fribourg (Switzerland) and Institut National des Télécommunication, Evry (France). Then, she joined the Department of Computer Science, University of Sheffield (UK) as Research Associate within the Speech and Hearing Research. In 2010 she joined the Department of Telecommunications, Networks and Computer Science at the National School of Applied Sciences, teaching engineering students in the area of software engineering with a focus on web/mobile app design and development. Her research interests include biometrics technologies, speech processing and issues related to BigData analytics.