Automatic Speech Recognition: An Improved Paradigm

Tudor-Sabin Topoleanu[1] , Gheorghe Leonte Mogan[1]

[1] B-dul Eroilor, Nr. 29, 500036, Brasov, Romania
{tudor.topoleanu, mogan}@unitbv.ro

Abstract. In this paper we present a short survey of automatic speech recognition systems underlining the current achievements and capabilities of current day solutions as well as their inherent limitations and shortcomings. In response to which we propose an improved paradigm and algorithm for building an automatic speech recognition system that actively adapts its recognition model in an unsupervised fashion by listening to continuous human speech. The paradigm relies on creating a semi-autonomous system that samples continuous human speech in order to record phonetic units. Then processes those phoneme sized samples to identify the degree of similarity of each sample that will allow the detection of the same phoneme across many samples. After a sufficiently large database of samples has been gathered the system clusters the samples based on their degree of similarity, creating a different cluster for each phoneme. After that the system trains one neural network for each cluster using the samples in that cluster. After a few iterations of sampling, processing, clustering and training the system should contain a neural network detector for each phoneme unit of the spoken language that the system has been exposed to, and be able to use these detectors to recognize phonemes from live speech. Finally we provide the structure and algorithms for this novel automatic speech recognition paradigm.

Keywords: automatic speech recognition, natural language processing, probabilistic language acquisition, unsupervised learning of speech

1 Introduction

Speech recognition is the process which transforms vocal sounds into the meaning of these sounds, turning spoken language into written language or symbolic knowledge, and it can be either human or automatic.

Human speech recognition turns spoken language into an internal symbolic representation in our minds, thus turning speech into meaning. The process of human speech recognition is based on sequentially recognizing phonetic units by taking advantage of multiple acoustic cues and then aligning them to obtain a word or sentence, this process happens in our subconscious mind without our constant attention [1].

Automatic speech recognition systems use the same principle of sequentially recognizing speech units from an audio signal based on recognition models that have

been pre-trained to recognize these speech units, and then inferring the most probable word that is described by the succession of recognized speech units [2-3].

One feature that is not yet possible with the latter is the autonomous acquisition of the knowledge needed to recognize speech. Automatic systems are manually trained using databases of speech sounds [4], while humans are not born with the gift of speech recognition, instead they acquire it independently, progressively and autonomously [5].

Looking into the process of human language acquisition it becomes clear that it is a probabilistic endeavor [6]. Therefore it is possible to tackle this problem in a computational manner in order to program machines to acquire speech in an unsupervised manner [7-8]. However language acquisition in humans is an incremental process that starts with acquiring the capacity to recognize speech and then progressing to the process of language learning which relies on that former capacity [9]. Current research also suggests a strong link between perception and production of speech as these two processes constantly influence each other [10].

In this paper we propose an algorithm and structure for an automatic speech recognition system that allows semi-autonomous acquisition of speech recognition. The structure of the article is as follows: chapter two describes the contribution of our system to sustainability, chapter three is a short survey of automatic speech recognition systems, chapter four describes the structure algorithm and process that we propose for achieving semi-autonomous acquisition of speech recognition, chapter five addresses our current results and chapter six summarizes and details further work that will complete our research.

2 Contribution to Sustainability

The capacity to autonomously acquire, adapt and manage the database of speech samples needed to train neural networks for detecting phonetic units is the core innovation of our proposal. This ability gives our solution a higher level of autonomy and hence sustainability compared to current automatic speech recognition solutions.

This innovation gives the speech recognition system the capacity to self-maintain and also to adapt its database according to the inputs it receives, while also allowing the system to acquire its own samples of the language it will evolve to recognize. The purpose is to create a recognition system that needs little intervention from a human operator in order to be trained by being able to manage and train itself by processing its audio input.

Our intention is to create a system that mimics the human capability of acquiring speech and is therefore a self-sustaining software system that acts as a voice to text interface for other software systems. In our case the motivation for this research comes from creating an autonomous voice interface for mobile robots that will become a component module of a control architecture for mobile robots. Following this key requirement of self sustainability we set out to design a system that will be capable of acquiring the skill of speech recognition. However we do not think that our proposed solution will be useful only for mobile robot applications, we hope that

it will find a use in other domains as well, for this reason we want to make it as easy as possible to exchange the recognition knowledge between instances of our system in order to allow other researchers to avoid, if so desired, the semi-autonomous acquisition phase.

3 A Short Survey of Automatic Speech Recognition

State of the art speech recognition systems can be split into a few categories: voice detection algorithms, user voice recognition, automatic speech recognition, emotion recognition and natural language processing.

Voice detection algorithms (VAD) simply detect when a recorded or live audio signal contains voice signals. One state of the art VAD system uses wavelet transforms in a wavelet filter bank for feature extraction from the input signal and then uses a Support Vector Machine (SVM) to train an optimized decision rule based on those extracted features [2, 11]. Another method that uses statistical models and machine learning for VAD employs generalized gamma distribution and learns from a speech database using minimum classification error (MCE) and SVM [12]. Another approach for robust VAD uses wavelet packet transform to analyze and extract transient components of speech and can extract speech activity from sources with a poor signal to noise ratio [13].

The current paradigm for automatic speech recognition consists of using either discriminative training models or generative training models [3]. Discriminative models based on Hidden Markov Models (HMMs) that are trained using a speech database and then used to recognize speech are a very important approach to realizing automatic speech recognition and could be extended to every corner of recognizer design [14]. Another discriminative approach to automatic speech recognition is based on neural networks which can be used for recognizing phonetic units, syllables or words [15] or the meaning of natural language [16].

The problem of active learning for speech recognition has been tackled before [17-18] however the approach is somewhat different since it relies on statistical processing and annotated corpus for processing real input data and minimizing the uncertainties from within it.

A current common trend is to create speech recognition systems that integrate multiple phonetic and acoustic feature extraction methods with language level modeling or language processing to reduce recognition errors and increase robustness of the system [19-22]. This is helpful because it provides multiple ways of detecting and eliminating recognition errors.

The common limitations and shortcomings of speech recognition systems is the requirement of using database of speech sounds for manually training the recognition models. The lack of support for less common languages, due to insufficient resources for compiling speech databases, and the high level of knowledge and technical skill required to train such recognition models and to create ASR systems.

4 An Improved Paradigm

Our proposed structure uses three levels for storing speech samples. The first level is a temporary buffer which has the function of storing all recorded speech samples until a limit has been reached. This limit is a parameter of our system (L1), once the limit is reached all samples from the buffer are analyzed and the useful ones are transferred to the second level of storage for further processing while the un-useful samples are removed, hence the buffer is now empty and ready for a new iteration.

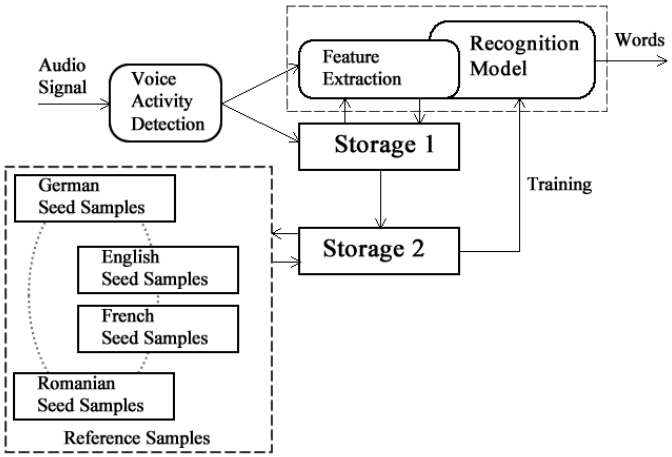

Fig. 1. The general structure of our system showcasing the main modules and data flow between them and the seed samples needed for clustering into phonetic clusters. As well as the separation between the recognition process, the acquisition process and the reference samples needed for clustering.

The second level of storage has the purpose of storing the samples selected from the temporary buffer and clustering them according to how similar they are. This level also has a limit of samples that can be stored, this is the second system parameter (L2), when the limit is reached clustering of samples is initiated at the end of each clustering process the last half of samples in each cluster are deleted. The selection of samples is made using a genetic algorithm that evaluates the fitness of each one and keeps only the best half of samples in each cluster.

The third and final level contains the neural networks that are trained with the samples in each cluster after the halving deletion. This level contains the models that are actually used to recognize phonetic units from live speech and therefore this level contains the recognition knowledge that can be transferred between instances of our system. This knowledge export/import feature is necessary since we consider that it would simplify the process of testing and evaluation before implementing a fully functional recognition system on a mobile robot, rather than testing and evaluating the system directly on the mobile robot.

Each neural network detector has the task of identifying a phonetic unit and is linked to a node in a HMM. In order to resolve the problem of the correct identification and clustering of phonetic units for each detector cluster the application will have to make the connection between phonetic unit and the equivalent written phoneme. In order to solve this problem we provide an "innate" set of samples representative of every phonetic unit of the language considered that the system contains from the start, these will be called reference or seed samples (these can be samples used in existing speech databases, or made especially for this task by requesting users to speak a pre-defined paragraph in their) and their analysis enables accurately clustering the recorded samples based on evolutionary feature similarity between the recorded samples and the seed samples. The number of reference samples is fixed and does not change, this being another system parameter (S).

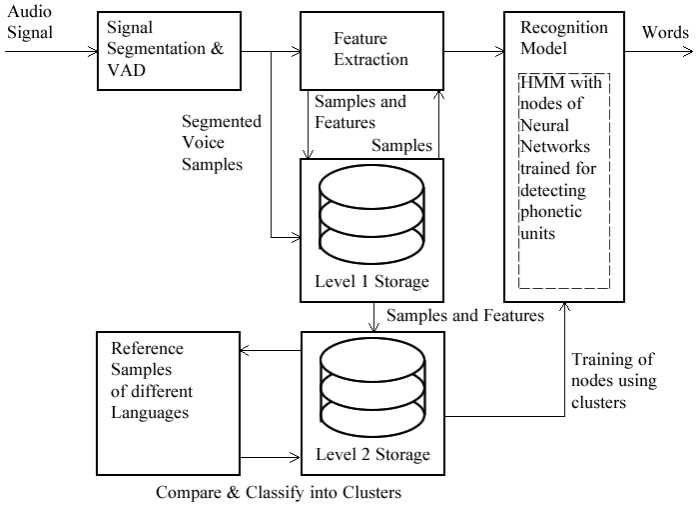

Fig. 2. Detailed view of the proposed structure.

For feature extraction we will use computationally fast features currently employed in state of the art systems (wavelet transform, mel-frequency cepstrum coefficients) as well as a slower process that analyzes the distinctive features (consonant, sonorant, syllabic, laryngeal, manner and place features of phonetic units) of each sample beginning with the seed samples and then with the recorded samples in order to have a wide coverage of features for any given speech sample. To allow for such complexity we will have to use one neural network for each phonetic unit of the language. Clustering of recorded samples needed to train each detector will be made using a similarity determination algorithm based on a genetic algorithm that will compare the features of each recorded sound to those of each reference sample and classify the sample as belonging to the cluster of the most similar phonetic unit.

The fast feature extraction methods are used to obtain features from live speech which are then sent to the detectors that have also been trained with these types of

features. The slow features are used for the clustering process, because they can't be a viable option when recognizing live speech since they are too slow to compute and because the clustering process has to rely on as much features as possible. When the storage limit is reached the feature extraction, clustering and training processes begin while the system stops recording and focuses on these computationally demanding operations.

Table 1. Proposed acquisition algorithm

```
Initialize L1, L2, S, IMAX parameters, i = 0
WHILE (i <= IMAX)
  IF (Voice Activity Detected)
    Record sample to Level 1 Storage
  ELSE
    WHILE (Level 1 storage <= L1)
        Extract features
        IF (Level 1 storage = L1)
                Extract features from remaining recorded samples
                Move recorded samples to L2
                Clean Level 1 storage and
                Exit L1 WHILE
    WHILE (Level 2 Storage <= L2)
        Evaluate fitness of each sample with EA
        Cluster samples using seed samples features and EA
                IF (Level 2 storage = L2)
                        Evaluate fitness of remaining samples
                        Delete bottom half of each cluster according to
                        fitness
                        Exit L2 WHILE
    FOR (each c cluster from Level 2 storage)
        FOR (each sample s from sample cluster k)
                Train Neural Detector K with fast features of sample s
```

In parallel with this acquisition algorithm there will be a standard feature extraction and recognition algorithm that uses the HMM detector network, this thread begins to run in parallel with the acquisition algorithm thread once the maximum number of iterations has been reached and has higher priority over the later when voice activity is detected.

5 Discussion of Results and Critical View

Our results consist of the proposed algorithm, structure and reference speech samples for the Romanian and English language as well as part of the implementation of the system using the Java programming language and for the hardware part we have two AKG professional microphones and a professional USB audio interface from M-Audio.

The presented algorithm and structure combined provide an improved paradigm for acquiring speech recognition, in an autonomous way, and a means to create new speech databases for training recognition models. We consider this as the starting

point of our research into cognitive speech recognition and language acquisition for mobile robots.

6 Conclusions and Further Work

In this paper we have described a new paradigm for an automatic speech recognition system that mimics the acquisition of speech recognition capabilities by employing a novel structure and algorithm. Our research is in its incipient stage and therefore there is significant amount of testing and evaluation that remains to be done with our system.

Testing will have to validate the efficient and coherent acquisition of recognition knowledge as well as validating the recognition performances and capabilities of our solution by acquiring the knowledge to recognize Romanian and English languages by beginning with seed samples for each language.

The described structure and algorithms might potentially be improved and an optimum version must be found within the limits of our described three level framework. One way we could achieve this would be to optimize our system using genetic algorithms for identifying the optimum algorithms, neural networks models and parameters for our system. Another research possibility would be to use associative neural networks and self-organizing maps for the first two levels of storage instead of databases, and also implementing different types of neural networks for the selection and clustering processes in order to obtain an entirely neural networked based recognition system. Again optimizing this completely neural based structure would be possible by using evolutionary methods.

After we succeed in finding the best possible structure, algorithms, parameters and optimum settings for them we will proceed to designing and implementing a system that is capable of language acquisition while relying on our proposed system for continuous speech recognition.

Acknowledgement

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/88/1.5/S/59321

References

1. Toscano, J.C., McMurray, B.: Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics, Cognitive Science 34, 434--464 (2010)

2. Dixon, P.R., Oonishi, T., Furui, S.: Harnessing graphics processors for the fast computation of acoustic likelihoods in speech recognition, Computer Speech and Language 243, 510--526, (2010)
3. Gales, M.J.F., Flego, F.: Discriminative classifiers with adaptive kernels for noise robust speech recognition, Computer Speech & Language 24, 648--662 (2010)
4. Jansen A., Niyogi, P.: Point process models for event-based speech recognition, Speech Communication 51, 1155--1168 (2009)
5. Chater, N.,Christiansen M.H: Language Acquisition meets Language Evolution, Cognitive Science 34, 1131--1157 (2010)
6. Hsu, A.S., Chater,N.: The Logical Problem of Language Acquisition, Cognitive Science 34, 971--1016 (2010)
7. Seitz, A.R., Protopapas, A., Tsushima, Y., Vlahou, E.L., Gori, S., Grossberg, S., Watanabe, T.: Unattended exposure to components of speech sounds yields same benefits as explicit auditory training , Cognition 115, 435--443, (2010)
8. Van der Velde, F., de Kamps, M.: Learning of control in a neural architecture of grounded language processing, Cognitive Systems Research 11, 93--107, (2010)
9. Lightfoot, D.: Language Acquisition and Language Change, Wiley Interdisciplinary Reviews: Cognitive Science 1, 677--684 (2010)
10. Casserly E.D., Pisoni D.B.: Speech Perception and Production, Wiley Interdisciplinary Reviews: Cognitive Science 1, 629--647 (2010)
11. Chen S-H., Guido, R.C., Truong, T-K., Chang Y.: Improved voice activity detection algorithm using wavelet and support vector machine, Computer Speech and Language 24, 531--543 (2010)
12. Shin, J.W., Joon-Hyuk Chang, J-H., Kim, N.S.: Voice activity detection based on statistical models and machine learning approaches, Computer Speech and Language 24, 515--530 (2010)
13. Mohadese Eshaghi, M.R. Karami Mollaei, Voice activity detection based on using wavelet packet, Digital Signal Processing 20, 1102--1115 (2010)
14. Jiang, H.: Discriminative training of HMMs for automatic speech recognition: A survey, Computer Speech and Language 24, 589--608 (2010)
15. Dede, G., Sazlı, M.H.: Speech recognition with artificial neural networks, Digital Signal Processing, 20, 763--768 (2010)
16. Majewski, M., Zurada, J.M.: Sentence recognition using artificial neural networks, Knowledge-Based Systems 21, 629--635 (2010)
17. Yu, D., Varadarajan, B., Deng, L., Acero, A.: Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion , Computer Speech and Language 24, 433--444 (2010)
18. Wu, W-L., Lu, R-Z., Duan, J-Y., Liu, H., Gao, F., Chen, Y-Q.: Spoken language understanding using weakly supervised learning, Computer Speech and Language 24, 358--382 (2010)
19. Siniscalchi, S.M, Lee, C-H.: A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition, Speech Communication
20. Nair, N.U., Sreenivas, T.V.: Joint evaluation of multiple speech patterns for speech recognition and training, Computer Speech and Language 24, 307--340 (2010)
21. Chien, J-T., Chueh, C-H.: Joint acoustic and language modeling for speech recognition, Speech Communication 52, 223--235 (2010)
51, 1139--1153 (2009)
22. Srinivasan, S., Wang, D.: Robust speech recognition by integrating speech separation and hypothesis testing, Speech Communication 52 , 72--81 (2010)