RESEARCH ARTICLE | SEPTEMBER 06 2017

Current trends in small vocabulary speech recognition for equipment control FREE

Nikolaos Doukas; Nikolaos G. Bardis

AIP Conf. Proc. 1872, 020029 (2017)
https://doi.org/10.1063/1.4996686

Current Trends in Small vocabulary speech recognition for equipment control

Nikolaos Doukas[1, a)] and Nikolaos G. Bardis[1, b)]

[1] *Assistant Professor, Hellenic Army Academy, Varis – Koropiou Avenue, 116 72 Vari, Greece*
[2] *Associate Professor, Hellenic Army Academy, Varis – Koropiou Avenue, 116 72 Vari, Greece*

[a)]Corresponding author: ndoukas@sse.gr
[b)]bardis@ieee.org

Abstract. Speech recognition systems allow human – machine communication to acquire an intuitive nature that approaches the simplicity of inter – human communication. Small vocabulary speech recognition is a subset of the overall speech recognition problem, where only a small number of words need to be recognized. Speaker independent small vocabulary recognition can find significant applications in field equipment used by military personnel. Such equipment may typically be controlled by a small number of commands that need to be given quickly and accurately, under conditions where delicate manual operations are difficult to achieve. This type of application could hence significantly benefit by the use of robust voice operated control components, as they would facilitate the interaction with their users and render it much more reliable in times of crisis. This paper presents current challenges involved in attaining efficient and robust small vocabulary speech recognition. These challenges concern feature selection, classification techniques, speaker diversity and noise effects. A state machine approach is presented that facilitates the voice guidance of different equipment in a variety of situations.

INTRODUCTION

Since the emergence of computing technology, it was envisaged by both the scientific community and the general public that it would be possible to construct human – computer interfaces that would be indistinguishable and equally intuitive nature as human to human interactions [1]. A fundamental element of such interfaces has always been Speech Recognition technology. Advances in the capabilities and operating speeds of processing equipment have enabled the design of speech recognition algorithms capable of reliable and robust recognition. However, the overwhelming advances in graphical user interface technology, dictated the usage of speech recognition as an alternative means of pressing a button [1]. Additionally, principal attention was given to high performance, computationally intensive applications that operate on service provider equipment, e.g. automated customer service systems that replace tone dialed menus.

The appearance of ubiquitous and cloud computing paradigms have given new directions in the related technologies, creating innovative types of requirements for speech recognition techniques. This change in direction was further accentuated by the development of processors with form factors suitable for providing limited but significant performance in embedded environments and the appearance of the necessity for reducing power consumption in computing equipment, an element of the green computing initiative [2]. Current speech recognition technology needs to operate in a seamless fashion, handing over between smart devices that may be interconnected in a mesh but do not necessarily have access to high performance networked or cloud computing resources. Hence, even though Apple's Siri and Google's Talk phonetic interfaces may give the impression that speech recognition technology has reached the absolute limit of performance equivalent to human, research in the field still presents a significant number of open issues. The problem may be redefined as enabling the user to interact with whatever application in a specific context, in a fashion similar to interacting with a capable human assistance [1]. This redefinition implies that virtually all current speech recognition solutions need to be revisited, reconsidered or even

Mathematical Methods and Computational Techniques in Science and Engineering
AIP Conf. Proc. 1872, 020029-1–020029-5; doi: 10.1063/1.4996686
Published by AIP Publishing. 978-0-7354-1552-2/$30.00

replaced so as to answer to current requirements. A more comprehensive approach to designing speech recognition solutions that enables operation in accordance with the above revised perspective therefore is not limited to the consideration of the inputs required by the system under consideration, but takes also into consideration the underlying process that is being interfaced via speech commands and the procedures required in order to have an intelligent, step – by – step interaction with this process. Elements of this paradigm that need to be addressed in purpose designed speech recognition algorithms include [1],

- Prediction and anticipation: Gradually develop a user specific model of the specific profile and behavior that will help service user needs, complete user interaction and even automate future transactions.
- Adaptivity: Algorithms dynamically modify their behavior according to evolving user or background conditions and fine – tune themselves according to user needs.
- Synergy: Algorithms produce output that may assessed together with corroborating information in order to arrive to the final decision
- Contextual information: Algorithms consider context, exploit particularities of the pursued interaction and proceed in steps in order to establish changes in the relevant context state variables.

The necessity for user – specific adaptation implies that speech recognition algorithms need to encompass elements of machine learning [3]. For effective performance, three different paradigms of machine learning need to be included in the algorithm design [1], [3]. The first one is supervised learning whereby an initial connection to the user is established and a baseline performance is achieved. Additional unsupervised learning may be achieved using contextual information, in order to infer the correctness of recognition. Finally, reinforced learning may be pursued by using spontaneous user reactions to recognition results, captured as further user phonetic input beyond the initially perceived speech period. Algorithms designed to pursue such training, especially unsupervised and reinforced training, are inevitably based on a number of heuristic inference rules. The task of minimizing or eliminating arbitrariness in the design of such heuristic rules, represents a significant challenge with respect to algorithmic design. Depending on the processing power available, successful machine learning algorithms attempt to discover and reproduce human mental operations in order to achieve enhanced performance. Such technological advances have been pursued by researchers in the area of Deep Learning [1].

The considerations outlined above demonstrate the necessity for the design of innovative speech recognition algorithms in the context of integrated, context aware, adaptive, human – machine interfaces. This study focuses on ultra – lightweight algorithms that provide solutions to a variety of problems arising from current technological necessities [1]. These necessities include the design of embedded equipment for

- safety devices
- conserving energy while maintaining the necessary performance
- facilitating data collection
- facilitating medical supervision
- enhancing connectivity of user equipment

This paper reviews enabling technologies for extracting features suitable for the task of producing context aware speech recognition decisions for the purpose of equipment control as defined earlier on. Relevant classification techniques are outlined, as presented in current state of the art literature that are capable of assessing the extracted features, given time and processing limitations. Techniques that have been proposed by researchers are listed for addressing problems arising due to speaker diversity effects. Particular attention is given to the effects of noise interference in the input. A suggested design of a state machine based limited vocabulary speech recognizer is presented that is capable of context aware decisions.

OUTLINE OF SPEECH RECOGNITION SYSTEM

In the previous section it was analyzed that the problem of speech recognition may be redefined in an application specific manner, as the problem of enabling the user to interact with whatever application in a specific context, in a fashion similar to interacting with a capable human assistance [1]. Additionally, a generic structure of a speech recognizer may be described that includes all the fundamental elements of processing elements necessary for robust recognition [4]. This structure is given in Figure 1. In this Figure, speech input is fed to a noise reduction component. This component may entail some classical filtering operations, such a low pass to eliminate excess noise, heuristic filtering such as notch filters to eliminate application specific interference and any other blind operations based on prior knowledge, that are considered to remove useless signal components.

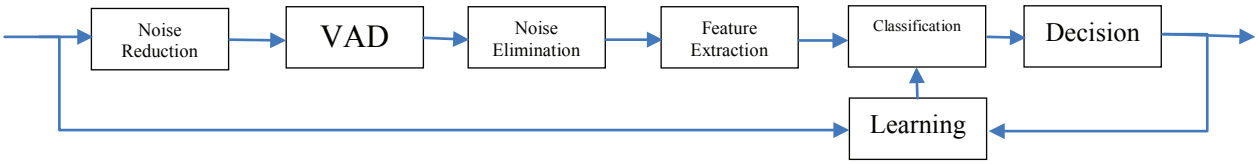

FIGURE 1. Block diagram of generic speech recognition system

The Voice Activity Detection (VAD) component is responsible for segmentation of the speech signal into single word segments. Essentially, since this study is dealing with isolated command recognition, this component is required to eliminate trigger the recognition process when a command is being issued. In other words it becomes more important in this case to produce an accurate estimate of the instant of the onset of speech, while errors in the detection of the ending instance of a command are more tolerable.

The Noise Elimination block processes the isolated signal segment and attempts to produce an initial decision on whether this segment is candidate to be a command, or is a speech – like noise burst. Documented signal characteristics and heuristic rules may be employed in order to make this classification.

The Feature Extraction block makes all measurements necessary on the selected segment that will be used for the final decision. The measured features may be used to update long term statistical measures to aid the adaptation of the process to varying background conditions.

The Classification block assesses the measures features and produces a set of final decision variables that are eventually used for the final decision. Based on the measurements, a segment may still be rejected at this stage as more elaborate interfering noise, or an irrelevant or unknown command. During the learning phase (whether supervised, unsupervised or reinforced), the Classification and Decision block parameters will be modified so as to reflect the application and user specific speech and background signal characteristics. In adaptive systems, these parameters may be under continuous modification so as to track variations in such characteristics.

Voice Activity Detection Principles

Baseline parameters for voice activity detection are Frame Energy and Frame Zero Crossing rate. The Frame Energy may straightforwardly calculated as:

$$FE = \sum_{i=1}^{N} x_i^2 \tag{1}$$

with x being the signal value and N the limit (end) of the determined speech period. Zero crossings may be measured using different principles, with a simple method being [4]:

$$ZC = \sum_{i=1}^{N} sign(x_{i-1} \cdot x_i) \tag{2}$$

where the sign function gives a value of 1 for negative values. Suitable thresholds may be used to determine combinations of values of FE and ZC for which frames are classified as being speech or noise. A simple variable that may be used for a thresholding decision may be calculated as:

$$T_{thr} = \frac{1}{N} \sum_{i=1}^{N} FE_i \cdot ZC_i \tag{3}$$

Using this decision variable, periods of high energy with a very small number of zero crossings may be distinguished from high energy periods with significant zero crossings. Additionally, high energy periods may be further classified according to their zero crossing rate, whereby a very high zero crossing rate may indicate strong noise and a medium zero crossing rate implies the existence of speech. Further, application dependent classifications are possible.

Features and Feature Extraction Techniques

In frames that have been initially classified as containing noise, calculations are initiated so as to extract the relevant features. Usual choices of features include the short term Fourier transform and the Mel Frequency Cepstral Coefficients (MFCC).

The Fourier analysis may reveal a variety of characteristics of an input signal, such as transient effects, long term periodic nature or spectral balance [4]. Voice signals will in general exhibit a complicated spectral structure that varies with time. Short term Fourier spectral analysis will hence provide an efficient way of acquiring a stable estimate of the spectral structure of the signal during a specific period, namely the period of the determined segment or frame. This is estimated as:

$$X_n(e^{j\omega}) = \sum_{m=1}^{N} x_n e^{-j\omega m} \tag{4}$$

where 1 and N are the limits of the window of the signal that has been classified as speech. Large windows produce small time resolutions and high frequency resolutions with large estimation errors in each frequency bin. The resolution and estimation errors may be adjusted as necessary by considering multiple frames. However, such averaging is limited by the actual available speech duration which, for the isolated command recognition, is by definition not very long. Suitable analysis may yield appropriate compromises for the possible sub-frame analysis schemes. Another factor to be considered here is the required algorithm onset sensitivity, as using multiple frames will incur longer initial trigger times and possibly larger endpoint errors. Equation 4 implicitly introduces a rectangular windowing function. Depending on the available processing power and time restrictions, a better windowing function, such as a Hamming window may be applied.

The Mel Frequency Cepstral Coefficients are a function of the spectral estimate of the speech signal that have been shown in literature to be able to assign more importance to the spectral features that the human auditory mechanism perceives as more significant. In other words, the MFCC represent a feature vector that is close to the feature vector measured by the human ear. Calculation of the MFCCs involves a series of bandpass filters that mimic the auditory system response and formulate a frequency band sequence. The produced spectra are scaled via a logarithmic function and inverse transformed via the DCT. The process consists of two steps

$$m(l) = \sum_{k=o(l)}^{h(l)} W_l(k) \left| X_n(k) \right|, l = 1,2,...L \tag{5}$$

and

$$c_{mfcc}(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} \log m(l) \cos\left[(l - \frac{1}{2}) \frac{i\pi}{L} \right] \tag{6}$$

Classification Techniques

The MFCC feature vector described above suffers from a very high time – frame sensitivity. If the time frames of the speech segments considered are not properly time aligned, the resulting MFCCs cannot be identified as being highly similar. A relatively simple and robust algorithm may be employed to rectify this problem. This algorithm is Dynamic Time Warping (DTW) [5]. DTW is used to determine an optimal alignment between speech utterances that have undergone non-linear time distortions. Additionally, DTW gives the advantage that all available training speech data may be used, thus giving maximum usage of the training and reliable recognition.

CONCLUSIONS

Speech processing principles were reviewed that enable the design of robust and accurate speech recognition schemes that use context information in order to achieve isolated command recognition. Current challenges involved in attaining efficient and robust small vocabulary speech recognition were presented. These challenges concern feature selection, classification techniques, speaker diversity and noise effects. A state machine approach will be presented that facilitates the voice guidance of different equipment in a variety of situations.

REFERENCES

1. H. Alkhatib, P. Faraboschi et al. IEEE CS 2022 Report. IEEE Computer Society, January 2017.
2. Definition - What does Green Computing mean? https://www.techopedia.com/definition/14753/green-computing, visited January 2017
3. Deng, Li, and Xiao Li. "Machine learning paradigms for speech recognition: An overview." *IEEE Transactions on Audio, Speech, and Language Processing* 21.5 (2013): 1060-1089.
4. Mi, Li, and Xie Quanying. "Research on speech recognition system in complex environments." BioTechnology: An Indian Journal 10.20 (2014).
5. Broekx, Lize, et al. "Comparing and combining classifiers for self-taught vocal interfaces." *SLPAT 2013* (2013): 21.
6.
7. M. P. Brown and K. Austin, Appl. Phys. Letters **85**, 2503–2504 (2004).
8. R. T. Wang, "Title of Chapter," in *Classic Physiques*, edited by R. B. Hamil (Publisher Name, Publisher City, 1999), pp. 212–213.
9. C. D. Smith and E. F. Jones, "Load-cycling in cubic press," in *Shock Compression of Condensed Matter-2001*, AIP Conference Proceedings 620, edited by M. D. Furnish *et al*. (American Institute of Physics, Melville, NY, 2002), pp. 651–654.
10. B. R. Jackson and T. Pitman, U.S. Patent No. 6,345,224 (8 July 2004)
11. D. L. Davids, "Recovery effects in binary aluminum alloys," Ph.D. thesis, Harvard University, 1998.
12. R. C. Mikkelson (private communication).