ISSN: 2229-6948(ONLINE)
DOI: 10.21917/ijct.2016.0201

ACOUSTIC SPEECH RECOGNITION FOR MARATHI LANGUAGE USING SPHINX

Aman Ankit[1], Sonu Kumar Mishra[2], Rinaz Shaikh[3], Chandraketu Kumar Gupta[4], Prakhar Mathur[5], Soudamini Pawar[6] and Anil Cherukuri[7]

Department of Computer Engineering, D Y Patil College of Engineering, India

E-mail: [1]aman.ankit1803@gmail.com, [2]sonukumar141@gmail.com, [3]shaikhrinaz94@gmail.com, [4]chanduaman96@gmail.com, [5]prakhar.jhl@gmail.com, [6]psoudamini@yahoo.co.in, [7]anilcherukuri@persistent.com

Abstract

Speech recognition or speech to text processing, is a process of recognizing human speech by the computer and converting into text. In speech recognition, transcripts are created by taking recordings of speech as audio and their text transcriptions. Speech based applications which include Natural Language Processing (NLP) techniques are popular and an active area of research. Input to such applications is in natural language and output is obtained in natural language. Speech recognition mostly revolves around three approaches namely Acoustic phonetic approach, Pattern recognition approach and Artificial intelligence approach. Creation of acoustic model requires a large database of speech and training algorithms. The output of an ASR system is recognition and translation of spoken language into text by computers and computerized devices. ASR today finds enormous application in tasks that require human machine interfaces like, voice dialing, and etc. Our key contribution in this paper is to create corpora for Marathi language and explore the use of Sphinx engine for automatic speech recognition.

Keywords:

Automatic Speech Recognition (ASR), Speech Analysis, Modeling Techniques, Acoustic-Phonetic Approach, Pattern Recognition Techniques, Language Modeling, Hidden Markov Model, Sphinx Engine, Phonemes, Lexicons

1. INTRODUCTION

Speech recognition technique is the process of mapping an acoustic waveform into a text (or the set of words) which should be equivalent to the information to be conveyed by the spoken word. In computer system domain it is defined as the ability of computer systems to accept input in the form of spoken words(audio format) and map into a text format in order to proceed with further computations [1].

It is one of the most important applications of Natural Language Processing (NLP) which is an active area of research and development. For NLP, a basic unit of speech recognition is the intermediate form of speech information. There are different categories of intermediate unit like- sub-phoneme units, phones with right or left context, biphones, diphones and variations, dyads or transemes, avents and etc. Comparatively, ASR for Indian languages is still at its infancy, hence, ongoing decade shows growing interest and huge scope in this field. The various functions involved are- Speech analysis, Feature extraction, Acoustic modelling Language and lexical modelling Recognition.

1.1 CLASSIFICATION OF SPEECH RECOGNITION SYSTEM

ASR can be classified in several different classes [2] based on speech utterance, speaker model and the vocabulary they recognize.

a) ***Speech Utterance Based:*** The vocalization (speaking) of words is an utterance. Its various types are: Isolated Words, Connected Words and Continuous Speech.

b) ***Speaker Model Based:*** The two main groups based on speaker models are speaker dependent and speaker independent. Speaker independent systems are difficult to develop, expensive and are less accurate than speaker dependent systems. But, these systems are more flexible and tolerant.

c) ***Vocabulary Based:*** Speech recognition systems can be defined on the basis of quantity of words stored in the dataset. Small vocabulary dataset contains tens of words, Medium vocabulary dataset contains hundreds of words, large vocabulary dataset contains thousands of words, and very large vocabulary dataset contains millions of words or Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word.

2. DIFFERENT APPROACHES TO AUTOMATIC SPEECH RECOGNITION

2.1 ACOUSTIC-PHONETIC APPROACH [3]

In Acoustic-Phonetic approach acoustic characteristics of speech sound is considered. In this approach smallest units of sound i.e. phonemes are found and provided with a label or tag [4]. Every language consists of a set of phonetic units which are combined together to give an entire word or sentence. All these units have a set of acoustics properties. Three techniques are mostly used for language identification. Gaussian mixture modeling [5], Problem phone recognition, and Support vector machine [6] classification.

2.2 PATTERN RECOGNITION APPROACH [7]

Pattern recognition approach (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993) is based on machine learning concepts which can use both labeled and unlabeled data. In this approach very sophisticated mathematical framework are designed which helps to establish uniform speech pattern representations which makes pattern comparison very reliable. It uses a set of labeled training samples involving formal training algorithms like clustering, classification and regression. Speech can be represented as a speech template or a statistical model (e.g., a Hidden Markov Model [8]) and can be applied to a sound (smaller than a word), a word, or a phrase. Template based, Dynamic time warping, Knowledge based, Statistical base and Stochastic are some of the different pattern recognition approaches.

2.3 ARTIFICIAL INTELLIGENCE (AI) APPROACH

This is the fastest and most sophisticated approach to speech recognition. It uses neural network techniques which are based on solutions derived for speech recognition. Significant progress has already been achieved in this area but the major concern is the robustness of the system. Several training algorithms are employed for real time response [9] [10].

3. SPEECH RECOGNITION STEPS

The Fig.1 shows the general model for acoustic speech recognition.

3.1 INPUT SPEECH

Continuous spoken sentences or voice stored in some input device can be used as input speech source. The ASR process uses comparison based technique where each and every utterance of words is stored in a vocabulary table. The spoken words are then analyzed properly and converted into a particular bit pattern that is stored in the vocabulary table.

3.2 ACOUSTIC MODEL

Acoustic modeling is basically based on the statistical representations of each of the distinct sounds that makes up a word. These statistical representations are assigned labels or tags called as phonemes. An acoustic model is created by using a large database of speech (also called a speech corpus) and using several special training algorithms to create statistical representations for each phoneme in a language. These statistical representations are defined as Hidden Markov Models (HMMs). Each and every phoneme has its own HMM. Sphinx tool makes use of HMMs for acoustic modeling of feature vectors.

3.3 FEATURE EXTRACTION [11] [12]

In this step, physical effects of input signal are parameterized. The essential attributes of speech are extracted whereas unimportant characteristics or noises are reduced. It includes various techniques like:

- Linear predictive analysis (LPC)
- Linear predictive cepstral coefficients (LPCC)
- Perceptual linear predictive coefficients (PLP)
- Mel-frequency cepstral coefficients (MFCC) [13] etc.

Characteristic features of the sound wave like frequency, amplitude and tone is extracted in feature extraction and stored in suitable format. The main aim is to reduce the redundancies present in the speech signal. Features are extracted based on frames (windows).

3.4 LANGUAGE AND LEXICAL MODELING [14]

Language modeling technique is based on assignment of probabilities to different word sequences. Language model enables the system to conspicuously choose between words that are similar. N-gram language models used in ASR systems are based on statistical approach and are used to search for correct word sequences by predicting the likelihood of nth word on the basis of *n*-1 preceding words. Language modeling can be either deterministic (Grammar based) [15] or Stochastic (Statistical).

In lexical modeling there is phonetization process. The speech is represented as lexemes using phonemes. The Speech recognition systems contain internal lexicons which specify spoken or recognizable words in a language. It also predicts how a word is expected to be pronounced. Phonemes describe continuous words spoken and represent valid set of tokens. These tokens are used to define the *n*-numbers of pronunciations of the same words using phonetic spellings.

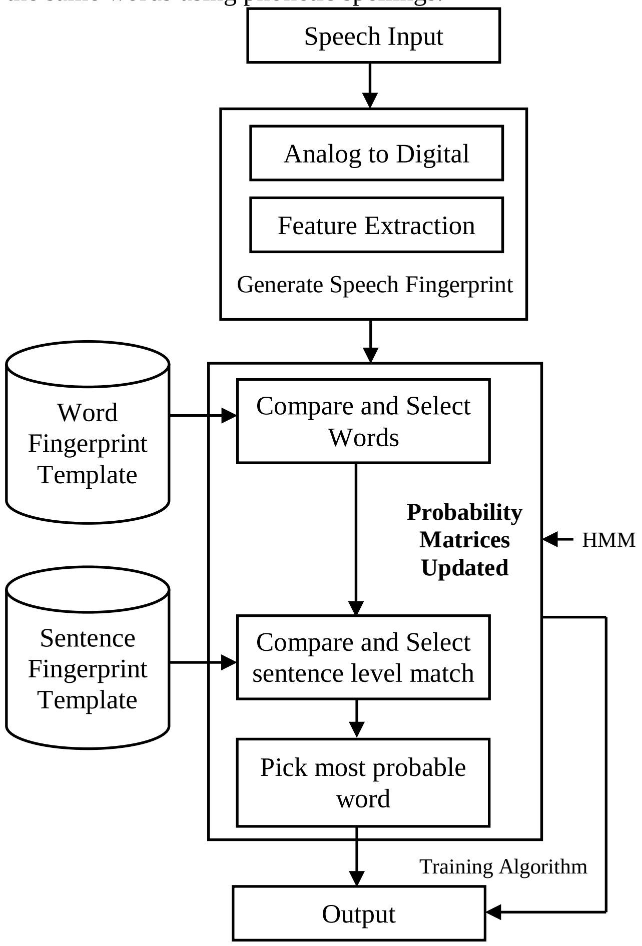
Speech Input
Analog to Digital
Feature Extraction
Generate Speech Fingerprint
Word Fingerprint Template
Compare and Select Words
Probability Matrices Updated
HMM
Sentence Fingerprint Template
Compare and Select sentence level match
Pick most probable word
Training Algorithm
Output

Fig.1. General Acoustic Speech Recognition Model

3.5 OUTPUT GENERATION

The output of an ASR system is the recognition and translation of spoken language into text by computers and computerized devices.

The output obtained is cast into a number of applications widely used today like, voice dialing domotic appliance control and etc.

4. TOOLS FOR ASR

The following Table.1 enlists various tools used for speech recognition and translation in the proposed system.

JSAPI is an open source Java API with cross-platform support for command and control recognizers, speech synthesizers and dictation systems. Several implementations can be created by third parties.

Table.1 Various Tools used for ASR

Tools	Usage
JSAPI	Java Speech API (JSAPI) is an application programming interface providing cross platform support for speech synthesizers and is used for acoustic modeling. It supports speech synthesis and speech recognition.
Sphinx	Sphinx 4 is a latest version of Sphinx series of speech recognizer tools, written completely in Java programming language. It provides a more flexible framework for research in speech recognition.
LMtool	LM Toolkit is a set of tools used for language modeling work in research. It is used for converting Marathi words into phonemes.
Google Translate API	The Google Translate is used to dynamically translate text between a numbers of language pairs and allows websites and programs to integrate with Google Translate using programming.
Smart tools	This consists of a set of applications for android and is used for noise level detection.

Sphinx library is based on Hidden Markov Models (HMMs) and an n-gram statistical model for converting speech in textual format. It is an open source API written in Java. It has a very flexible framework and support for research in the field of Speech Recognition for several languages.

5. PROPOSED SYSTEM

Sphinx4 library [16] is a research based library for speech recognition. The proposed system uses this for speech to text conversion in Marathi language. The main focus of the research has been the analysis of phonetic structure of words spoken in Marathi language. The Fig.2 CMU-Sphinx Architecture, the main blocks are Front End, Decoder and Linguist.

a) ***Front End:*** Digital Signal Processing (DSP) is performed on input speech signal i.e. feature extraction.

b) ***Linguist:*** It is also known as knowledge base which is needed by decoder. It consists of three modules:

c) ***Acoustic Model:*** The statistical representation of a sound.

d) ***Dictionary:*** Represents how a word is pronounced. The CMU-Sphinx provides a very flexible dictionary. It can store n-numbers of phonetic structure of the same word; this feature can be used to reduce the conflict which arises due different pronunciation of the same word mainly due to regional differences.

e) ***Language Model:*** It shows the probability of occurrence of a word.

f) ***Search Graph:*** The graphical structure produced by linguist using knowledge base.

g) ***Decoder:*** It is the main block in CMU-Sphinx. It reads the extracted features from front end, combines this with data from knowledge base and performs search to find most probable sequence of words.

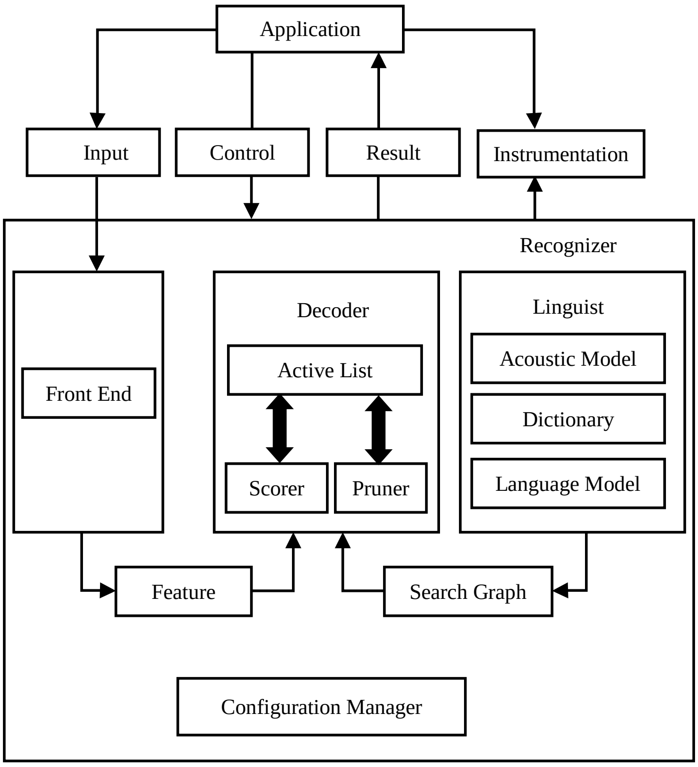

Fig.2. CMU-Sphinx Architecture

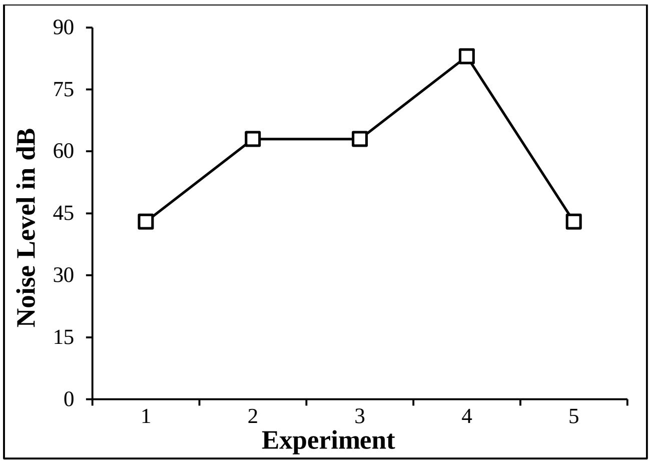

Fig.3. Experiment Graph under Different Noise Levels

6. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed system is designed with a Marathi dictionary of 10,000 words. The phonetic structure of words has been analyzed and stored in the dictionary format. A rudimentary speech recognition system recognizes the most commonly used Marathi sentences according to the JSGF (Java Speech Grammar Format) and words stored in the dictionary. Noise level has been recorded during the experiment with several speakers. The Fig.3

shows the experiment graph based on different noise levels. The system performance is then measured using the following metrics. The following table shows the results for our work.

The results are based on two parameters: Word Error Rate (*WER*) and Frequency.

$$WER = (S + I + D) / N \quad (1)$$

where, S = Substitution, I = Insertion, D = Deletion and N = Number of Words uttered.

$$\text{Frequency} = \frac{\text{number of times recognition is successful}}{\text{number of times speech is input}} \quad (2)$$

Table.2. Analysis of Word Error Rate with Speaker Variation

Speaker	Gender	Noise Level	WER	Frequency
Speaker-1	Female	40-45 dB	15%	90%
		60-65 dB	25%	70%
		80-85 dB	25%	60%
Speaker-2	Female	40-45 dB	30%	80%
		60-65 dB	40%	60%
		80-85 dB	40%	60%
Speaker-3	Female	40-45 dB	15%	100%
		60-65 dB	15%	70%
		80-85 dB	15%	70%
Speaker-4	Female	40-45 dB	20%	90%
		60-65 dB	25%	80%
		80-85 dB	25%	80%
Speaker-5	Male	40-45 dB	15%	100%
		60-65 dB	25%	80%
		80-85 dB	25%	70%
Speaker-6	Male	40-45 dB	20%	90%
		60-65 dB	20%	80%
		80-85 dB	25%	80%
Speaker-7	Male	40-45 dB	15%	80%
		60-65 dB	30%	60%
		80-85 dB	30%	60%
Speaker-8	Male	40-45 dB	20%	100%
		60-65 dB	30%	80%
		80-85 dB	35%	70%

The Table.2 gives a basic idea to evaluate the effectiveness of the proposed system. The Word Error rate and frequency is calculated for sentences comprising 4-6 words for different speakers for 25 sentences (limited grammar). It is observed that between the noise levels 40-45 dB, the proposed system performs better for most speakers with lower WER and higher frequency. But as noise level increases, the performance deteriorates. However, the interesting fact that is observed is the WER and frequency tend to be similar in relatively loud environment i.e. 60-65 dB and 80-85 dB. A frequency of 90% and above is at par for practical usage which has been achieved by the system under 50 dB of noise.

From the above results and Fig.4, it can also be stated that if we increase the support for language grammar, the efficiency would decrease slightly considering the words which have similar phonemes (i.e. spoken similarly).

In order to determine the accuracy of the system at different noise level, a line graph is shown as Fig.5. The percentage accuracy is calculated as the product of frequency and word correct rate (i.e. 1 - WER). A threshold is defined at 50% (in red color) and the performance above the threshold is considered acceptable for practical usage.

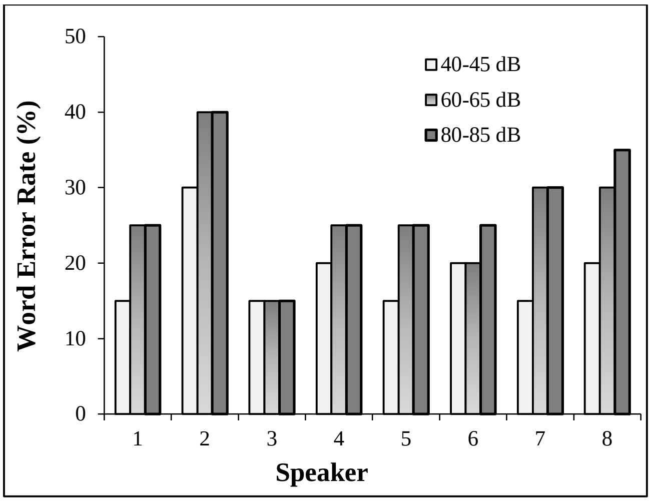

Fig.4. Result Graph Based On WER and Frequency

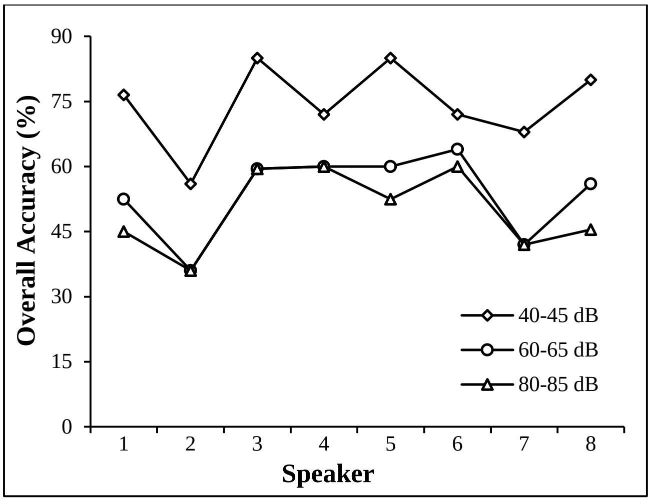

Fig.5. Accuracy graph illustrating system performance

7. CONCLUSION

Applications based on speech recognition are becoming popular as they can be for interacting with systems in local languages. In this paper we discussed speech recognition using CMU-Sphinx library. The library provides suitable framework for speech to text conversion and results can be improved by analyzing the dialects of Marathi language. We have briefly discussed the Speech Recognition System and various approaches which can be used to develop ASR in various languages. Hidden Markov Model is popular for building rudimentary ASR. With our work we intend to open gates for budding researchers in this enormous field of research.

8. FUTURE SCOPE

The acoustic speech recognition for Indian languages is a vast area of research as there is a lack of labeled speech dataset. We experienced the same while working for the Marathi language. There are certain challenges like effects of local dialect on the language, variation of accent, etc. On a better note of achieving high efficiency, experiment to gather phonetic structure of same word spoken by different type of people and working on DNNs can further improve results. Collecting data and labeling it, as suggested by the Woefzela tool [17] which is client-server based speech data gathering tool can prove very useful and pave a way to designing efficient and reliable systems.

REFERENCES

[1] Bassam A.Q. Al-Qatab and Raja N. Ainon, “Arabic Speech Recognition using Hidden Markov Model Toolkit (HTK)”, *Proceedings of International Symposium in Information Technology*, Vol. 2, pp. 557-562, 2010.

[2] B. Atal and L. Rabiner, “A Pattern Recognition Approach to Voiced Unvoiced-Silence Classification with Applications to Speech Recognition”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 24, No. 3, pp. 201-212, 1976.

[3] Tanja Schultz and Alex Waibel, “Language-Independent and Language Adaptive Acoustic Modeling for Speech Recognition”, *Speech Communication*, Vol. 35, No. 1, pp. 31-51, 2001.

[4] Amit Juneja and Carol Espy-Wilson, “Segmentation of Continuous Speech using Acoustic-Phonetic Parameters and Statistical Learning”, *Proceedings of the 9th International Conference on Neural Information Processing*, Vol. 2, pp. 726-730, 2002.

[5] Jean Luc Gauvain and Chin Hui Lee, “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, 1994.

[6] A. Ganapathiraju, J.E. Hamaker and J. Picone, “Applications of Support Vector Machines to Speech Recognition”, *IEEE Transactions on Signal Processing*, Vol. 52, No. 8, pp. 2348-2355, 2004.

[7] Dabbala Rajagopal Reddy, “Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave”, *The Journal of the Acoustical Society of America*, Vol. 40, No. 5, pp. 1273-1273, 1966.

[8] S.R. Eddy, “Hidden Markov Models”, *Current Opinion in Structural Biology*, Vol. 6, No. 3, pp. 361-365, 1996.

[9] Roger K. Moore, “Twenty Things We Still Dont Know About Speech”, *Proceedings of Workshop on Progress and Prospects of Speech Research and Technology*, pp. 1-9, 1994.

[10] Lalit R. Bahl, Fredrick Jelinek and Robert L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 2, pp. 179-190, 1983.

[11] S. Chu, S. Narayanan and C.J. Kuo, “Environmental Sound Recognition with Time-Frequency Audio Features”, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 6, pp. 1142-1158, 2009.

[12] Bishnu Prasad Das, “Recognition of Isolated Words using Features Based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers”, Ph.D Thesis, Jadavpur University, 2012.

[13] C.O. Dumitru and I. Gavat, “A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language”, *Proceedings of 48th IEEE International Symposium on Multimedia Signal Processing and Communications*, pp. 115-118, 2006.

[14] Helmer Strik and Catia Cucchiarini, “Modeling Pronunciation Variation for ASR: A Survey of the Literature”, *Speech Communication*, Vol. 29, No. 2-4, pp. 225-246, 1999.

[15] Daniel Jurafsky and James H. Martin, “*Speech and Language Processing*”, 2nd Edition, Prentice Hall, 1998.

[16] Hussein Hyassat and Raed Abu Zitar, “Arabic Speech Recognition using Sphinx Engine”, *International Journal of Speech Technology*, Vol. 9, No. 3, pp. 133-150, 2008.

[17] Nic J. De Vries, Jaco Badenhorst, Marelie H. Davel, Etienne Barnard, and Alta De Waal, “Woefzela-An Open-Source Platform for ASR Data Collection in the Developing World”, *Proceedings of 12th Annual Conference of the International Speech Communication Association*, pp. 3177-3180, 2011.