Radio-browsing for Developmental Monitoring in Uganda

Raghav Menon[1], *Armin Saeb*[1], *Hugh Cameron*[2],
William Kibira[2], *John Quinn*[2], *Thomas Niesler*[1]

[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]UN Global Pulse, Kampala, Uganda

rmenon@sun.ac.za, arsaeb@sun.ac.za, hcameron@cit.ac.ug,
williamkibira@gmail.com, john.quinn@one.un.org, trn@sun.ac.za

Abstract

We consider the extraction of information from broadcast radio speech in Uganda for the purposes of informing relief and development programmes by the United Nations. Although internet penetration in Uganda is low, mobile phones are ubiquitous and have made radio a vibrant medium for interactive public discussion. Vulnerable groups make use of radio to discuss issues related to, for example, agriculture, health, governance and gender by means of phone-in or text-in talk shows. We have compiled corpora and developed a radio-browsing system for Ugandan English and for two indigenous languages, Luganda and Acholi. The systems employ automatic speech recognisers using HMM/GMM, SGMM and DNN/HMM acoustic models as keyword spotters. We present the first results indicating promising performance of the radio-browsing system.

Index Terms: radio-browsing, radio broadcasts, keyword spotting, accented English, Luganda, Acholi

1. Introduction

There is recurrent anecdotal evidence that local African radio broadcasts often contain relevant, actionable information for development and relief programmes that are in advance of national level reporting (mainly television and print media) while also falling below their threshold for publication. This paper reports on the first steps in the development of a system to enable automated extraction of such information from radio broadcasts, with the aim of passing it to programme managers for verification and action, say within 24 hours of original broadcast.

In countries with prevalent internet access, social media communication is used as a tool by the population to voice concerns and views on various issues concerning the society, such as health, safety and food security [1, 2, 3]. Consequently, social media communication has recently been used effectively as a tool for early warning, monitoring and evaluation of development projects [4, 5, 6, 7]. In rural Uganda, however, low penetration of the internet and a predilection for graphic and textual content combine with a strong oral tradition to make radio by far the preferred medium of social communication. This is supported by the recent rise to ubiquity of mobile phones, providing access to phone-in or text-in talk shows. The monitoring of discussions on local and community radio, and the identification of keywords of interest, such as those referring to agriculture, health, governance and gender, make it a better alternative to online social media, particularly in rural areas. Through such radio-browsing, the needs, concerns and opinions of vulnerable groups can therefore be followed and assessed. We report first results for such a radio-browsing system.

Section 2 describes the system configuration, Section 3 the data compilation process and Section 4 gives a brief overview of the radio-browsing system. Section 5 describes the conducted experiments and discusses the results. Section 6 concludes our paper.

2. System Configuration

A block diagram indicating the major components in our radio-browsing system is shown in Figure 1. The radio stations host phone-in shows where the listeners call in and discuss issues of importance to them such as violence towards women, floods, malaria, adolescent pregnancy, price fluctuations, cholera etc. These discussions are transmitted by the radio stations. Radio broadcasts were recorded using an RTL2832U based DVB-T receiver, a Raspberry-pi and the GNU-Radio open source software tools. This setup allows multiple radio stations to be captured using a single device. The hardware encodes the captured audio as 8-bit PCM. The captured data passes through a keyword spotting system which searches for words of interest.

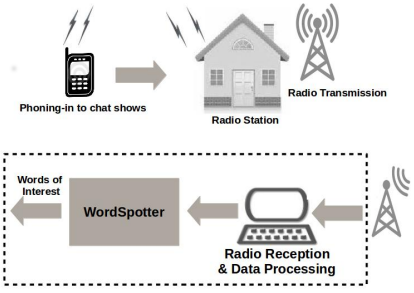

Figure 1: *System Configuration.*

Identification of keywords in spoken discussions can be performed using a keyword spotter, which is an application of automatic speech recognition (ASR). A keyword spotting system analyses a stream of speech audio to determine the occurrence of a specific set of words or phrases. The word or words can be in the form of audio or text. In the former case, it is usual to refer to a spoken term detection system. We will focus on the latter case, where search terms are provided in text form. In this paper keyword spotting is applied to the heavily accented and under-resourced variety of English spoken in Uganda and also to two indigenous Ugandan languages, Luganda and Acholi, to render the overall system multilingual. This will allow it to be applied widely in Uganda, since the relevant content is not always

in English. Rather it is frequently broadcast in local African languages whose speakers put forth viewpoints they would not express in English.

3. Data Compilation

In 2014, there were 216 registered FM radio stations across Uganda, broadcasting on 299 different frequencies. Figure 2 indicates the geographic distribution of FM transmitters using information provided by the Ugandan Communication Commission (UCC). Initially, broadcasts in the Kampala area of Uganda were targeted. However, in January 2016, additional recording posts were set up in Gulu and Moroto, which are located in the north and north-east of Uganda respectively, and where conflict, poverty and instability have been recurrent. Additional posts are being added in western, north-western and south-western Uganda (Kabarole, Arua and Mbarara) to cover the other areas where relief and development programmes are concentrated. All recorded data can be described as conversational and therefore challenging.

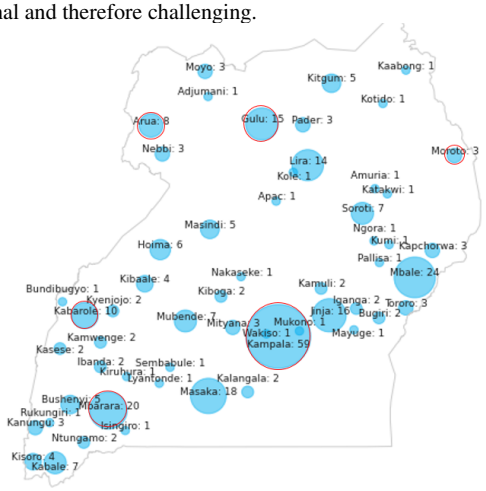

Figure 2: *Number of FM transmitters in Uganda by district and recording site.*

The recorded audio was annotated by mother-tongue speakers using PRAAT [8]. Silences, filled pauses and speaker noises are indicated in the transcripts. The small amount of Ugandan English gathered was augmented by a South African (SA) broadcast news data corpus [9]. In initial experiments, the SA data was found to be substantially closer to the Ugandan accent than alternative English sources. The data was divided into training and testing partitions as indicated in Table 1.

Language modelling data for English consisted of approximately 109 million words of text collected from major South African newspapers between January 2000 and March 2005 [9]. However, the available text for Luganda and Acholi was severely limited. Approximately 1,000,000 words obtained from the Bukedde newspaper were used for Luganda. In addition, the transcripts of the acoustic training sets were used. For each language, trigram language models were trained using the SRILM toolkit [10].

All pronunciation dictionaries were developed by phonetic experts, although the expertise available for Luganda and Acholi was much more limited than that available for English [11]. The databases available for system training in Ugandan English, Luganda and Acholi are small in comparison to those used by mainstream systems such as those developed for the Wall Street Journal (WSJ) [12, 13] and BABEL databases [14, 15]. This limitation presents difficult challenges.

Table 1: *Composition of training and testing corpora.*

	English		Luganda		Acholi	
	Train	Test	Train	Test	Train	Test
Utterances	12966	490	8172	794	4863	184
Speech	23h	37m	9h	62m	9h	18m
Speakers	758	48	380	75	203	NA

For wordspotting experiments, due to the constraint in the amount of testing data, we have selected 14 keywords which reflect the developmental needs in rural Uganda in all the three languages under consideration. System combination or fusion has been found by some authors to lead to improved wordspotting performance [16, 17, 14, 18] and has been evaluated in order to address the limitations on the data available.

4. Keyword Spotting System

Keyword spotting can be performed in various ways, as discussed in detail in [13, 19, 20]. Broadly one can distinguish between supervised and unsupervised approaches. Among supervised approaches one can further distinguish between approaches based on i) acoustic keyword models ii) large vocabulary continuous speech recognition (LVCSR) iii) subword models iv) query-by-example and v) events. Our keyword spotting system, illustrated in Figure 3, uses LVCSR to produce lattices, which are subsequently indexed and searched.

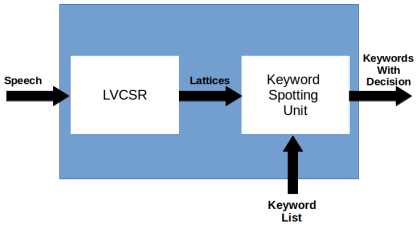

Figure 3: *Keyword Spotting System.*

Since the database that is being used is highly accented, under-resourced and under-represented, it is assumed that all the search words are known in advance. In this way we avoid the additional challenges of effectively dealing with out-of-vocabulary (OOV) words. A brief description of the components of Figure 3 is given in the following paragraphs.

4.1. LVCSR

The Kaldi speech recognition toolkit has been used for large vocabulary continuous speech recognition (LVCSR) [21]. This decoder is based on finite state transducers (FSTs) and incorporates the language model (LM), the pronunciation dictionary (lexicon) and context dependency into a single decoding graph [22]. Lattices are created for each test utterance. We have considered HMM/GMM, SGMM [23] and DNN/HMM [24] based acoustic models, as shown in Figure 4.

4.2. Keyword Spotting Unit

The keyword spotting unit operates on the lattices generated by the LVCSR system. First, suitable candidates are extracted from

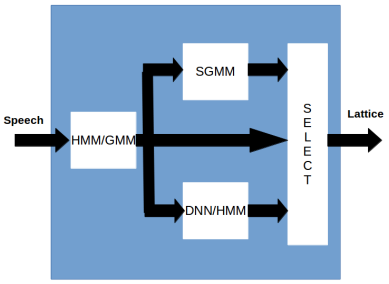

Figure 4: *LVCSR Unit.*

the lattices for detection. These candidates, together with posterior probabilities and start and end times, are stored in a reverse index. Next, a detector locates keywords in the index. Finally, a decision is made regarding the hypothesised keyword by comparing the posterior probability to a predetermined threshold [25] which is optimised for the application. Since the purpose of this set up is to address the concerns of vulnerable groups, it is assumed that the words of interest are predefined.

5. Experiments and Discussion

For lattice generation, HMM/GMM, SGMM, SGMM-BMMI and DNN/HMM acoustic models were trained and tested separately for Ugandan English, Luganda and Acholi using the corpora described in Table 1. The word error rates (WER) of these acoustic models are shown in Table 2. As it is shown, the best results are exhibited by DNN/HMM and SGMM-BMMI models. Lattices generated using these acoustic models were used in the keyword spotting experiments that follow.

Table 2: *WER for English, Luganda and Acholi.*

System	%WER (English)	%WER (Luganda)	%WER (Acholi)
HMM/GMM	46.71	57.15	59.04
SGMM	43.32	54.36	58.89
SGMM-BMMI	40.39	52.47	57.50
DNN/HMM	38.88	53.54	57.75

Keyword spotting performance is evaluated using Detection Error Tradeoff (DET) curves and also the NIST oracle measures, Actual Term Weighted value (ATWV) and Maximum Term Weighted Value (MTWV). Term Weighted Value (TWV) specifies a trade-off between the probability of a miss (P_{Miss}) and the probability of a false alarm (P_{Fa}). ATWV is the TWV at a particular threshold. Higher ATWV values indicate better keyword spotting performance. MTWV is the maximum ATWV over all thresholds. The threshold for the wordspotter has to be chosen so that it is located at the point of maximum ATWV (i.e. MTWV).

The oracle measures and the DET curves, which describe the system performance of a keyword spotter (KWS), are shown for Ugandan English, Luganda and Acholi in Table 3 and Figure 5 to Figure 7 respectively. Each of these figures show the DET curve of the best individual DNN/HMM system from Table 3 as well as the combination of all the individual systems. Table 3 shows that the ATWV and the MTWV values are not the same. This is due to the fact that the threshold which is being used and the actual threshold at which the best performance is obtained are different. The keyword spotting performance of the four systems improves in approximate sympathy with the WERs reported in Table 2, with the DNN system faring best. Possible normalisation or modification of scores for better keyword detection were suggested in [14, 16, 26, 27]. Application of these techniques did not, however, lead to improvement in our case.

Table 3: *Keyword spotting performance for various systems & system combinations.*

	English		Luganda		Acholi	
	ATWV	MTWV	ATWV	MTWV	ATWV	MTWV
HMM/GMM	0.5169	0.5278	0.2199	0.3095	0.2852	0.4146
SGMM	0.5860	0.6188	0.2584	0.3523	0.2845	0.4455
SGMM-BMMI	0.6018	0.6188	0.2357	0.3300	0.3554	0.4690
DNN/HMM	0.6115	0.6225	0.2807	0.3617	0.5210	0.5773
Combination (Avg)	0.5958	0.6500	0.3195	0.3660	0.4771	0.4828
Combination (Geo Avg)	0.6018	0.6134	0.3743	0.3743	0.4604	0.5157
Combination (Har Avg)	0.5946	0.6134	0.3717	0.3723	0.4788	0.5157
Combination (Condorcet)	0.6151	0.6565	0.2607	0.3587	0.3180	0.5034

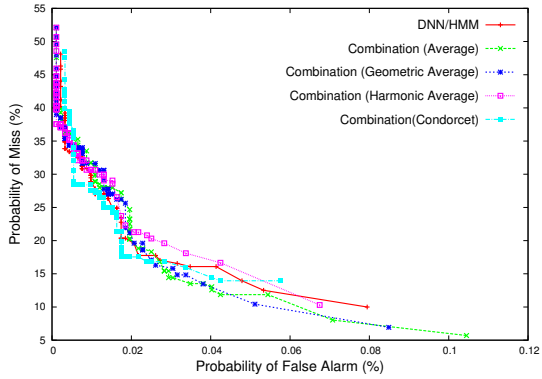

Figure 5: *DET curves for Ugandan English.*

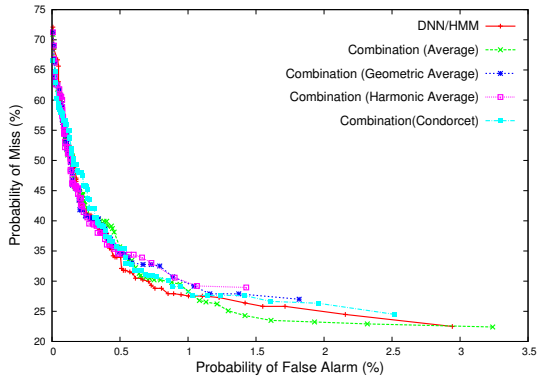

Figure 6: *DET curves for Luganda.*

In an attempt to improve performance, the outputs of the four systems were combined and results analysed. To do this outputs of the four systems were first aligned. Detections by different systems are considered to coincide when start and end times of the search term occur within 0.5 seconds of each other. If a particular system fails to hypothesize the search term within these start and end times, a score of zero is assumed. The scores of the four systems are combined using Equations 1-3, where N is the number of systems to be combined.

$$scoreA_f = \frac{1}{N}\sum_{n=1}^{N} score_n \quad (1)$$

$$scoreG_f = \exp\left\{\frac{1}{N}\sum_{n=1}^{N} ln(score_n)\right\} \quad (2)$$

$$scoreH_f = \frac{N}{\sum_{n=1}^{N}\frac{1}{score_n}} \quad (3)$$

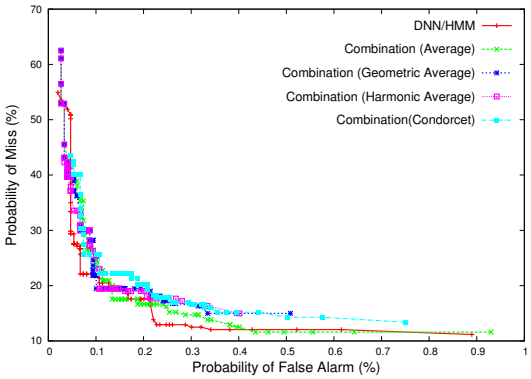

Figure 7: *DET curves for Acholi.*

$scoreA_f$, $scoreG_f$ and $scoreH_f$ are obtained by finding average, geometric average and harmonic average of the individual scores. The oracle measures for system combination, ATWV and MTWV, are given in Table 3. The results show that the ATWV (for the fixed threshold) and MTWV values are improved which is an indication that the system performance can be improved using system combination. The choice of the threshold plays an important role.

To test whether a non-linear combination would help in improving the system performance the Condorcet voting principle was used to choose between the systems. The scores assigned by the individual systems for each keyword are compared with each other and the total pairwise win for each system calculated. The system with the maximum pairwise win is chosen as the winner and the score by this system is taken as the final score. From Table 3, the Condorcet voting principle is seen to give better performance for English and similar performance to other combination methods for the other languages. Analysing the data, we found that the lattices for Luganda and Acholi contained many incorrect alternative hypotheses which were used during voting and hence influenced the average performance. The Condorcet voting principle was also used to combine the result obtained using system combination (combination of combinations). This approach equalled or was close to the best combined result. From Table 3, it can be seen that there is no specific combination that can be chosen as the best overall for a radio-browsing system. Instead, particular combination methods work well for a particular languages.

To evaluate the real-world utility of the radio-browsing system, we applied it to a large quantity of unannotated radio data in Ugandan English, Luganda and Acholi. This allowed us to obtain examples of the context in which keywords of interest were mentioned, and to make some assessment of the utility of these examples for development and humanitarian purposes. Radio recordings were made in three locations in Uganda: Kampala, Gulu and Moroto. Wordspotting was carried out on over 5500 hours of audio including over 650 hours of Luganda and 450 hours of Acholi. The keyword probability of error observed during July and August 2016 was 50-60% for Luganda and 40-50% for Acholi. This was deemed sufficient for real-world utility by the staff of the UN. (The volumes of data precluded exhaustive measurement of keyword errors). While the keyword error was slightly better at 30-40% for Ugandan English, the fact that local community members talk more freely in their mother tongues makes the Luganda and Acholi output more useful.

Table 4: Sample contexts of detected keywords in radio recordings.

Keyword	Context of detection
English	
Flood	the residents were warned to vacate the areas that constantly flood when it rains but they say if they shift their sugar cane will be stolen
Malaria	most health centers in Agago district are now without anti malaria drugs
Health	Awac health center three in Gulu district has been operating without toilets
Luganda (Context translated to English)	
Akawuka (means insect but common usage means AIDS)	Me if I have a house maid I make sure I check her HIV status, if she is positive I chase her and get another one because I cannot risk leaving my baby with her.
Omukyala (Wife or woman)	Police in Buloopa in Kamuli District has arrested a wife (woman). It is said that she stabbed her husband in the neck and the main reason is yet to been known.
Ababundabunda (Refugees)	There has been a Cholera outbreak in the camp for refugees known as Eregu Bidi-Bidi Camp. This is because of the level of sanitation which is very poor.
Acholi (Context translated to English)	
Olong kot [Lightning]	Lightning has struck 20 cows to death in Guda Palwo village in Madopei sub county in Lamwo district.
Yat [Medicine]	We have a problem in Ongango sub county, you find that if they have brought medicine today, but tomorrow you go to the hospital you will hear that ((NGs)) has not signed.
Kwan [Education]	People are objecting the Universal University education since the government has already failed to run both the Universal Primary education and Universal Secondary Education projects.

Table 4 shows some examples of the context of resulting keyword detections. We find in particular that eyewitness accounts of problems, often reported on small community radio stations, are valuable. These reports may individually be concerned with issues too minor to be reported through the mainstream media, but collectively provide actionable insights into the well-being of the population.

The experimental production pipeline, from community radio broadcast to recording upload to word spotting to scoring for relevance and finally to context analysis, has a latency of 8-24 hours. This allows eyewitness reports to be flagged for action before the news reaches the national media. Current experiments show that, for selected cases involving specified keywords and radio stations, the latency can be reduced to as little as 10 minutes.

6. Conclusions

We present first results for a radio-browsing system applied to three languages: Ugandan English, Luganda and Acholi. To our knowledge this is the first time speech recognition systems have been developed for the latter two languages. The radio-browsing system was tested using four different speech recognition architectures and system combination. It was found that the best individual system performance was achieved by a DNN/HMM system. System combination was demonstrated to achieve an improvement when the correct threshold is chosen. Ultimately the purpose of the system is to provide managers of relief and development programmes of the United Nations with current information in order to inform their decisions. The radio-browsing system is actively being deployed in Uganda since it provides performance that is satisfactory for real-world utility.

7. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation GPU equipment used for this research.

8. References

[1] S. Vosoughi and D. Roy, "A Human-Machine Collaborative System for Identifying Rumors on Twitter," in *Proc. ICDMW*, Atlanta City, 2015, pp. 47–50.

[2] K. Wegrzyn-Wolska, L. Bougueroua, and G. Dziczkowski, "Social Media Analysis for e-Health and Medical Purposes," in *Proc. CASoN*, Salamanca, 2011, pp. 278–283.

[3] P. Burnap, G. Colombo, and J. Scourfield, "Machine Classification and Analysis of Suicide Related Communication on Twitter," in *Proc. 26th ACM Conference on Hypertext and Social Media*, Cyprus, 2015, pp. 75–84.

[4] G. P. P. Series, "Analyzing Attitudes Towards Contraception and Teenage Pregnancy using Social Data," *Global Pulse Project Series*, no. 8, 2014.

[5] ——, "Mining Citizen Feedback Data for Enhanced Local Government Decision-Making," *Global Pulse Project Series*, no. 16, 2015.

[6] ——, "Understanding Immunisation Awareness and Sentiment Through Social and Mainstream Media," *Global Pulse Project Series*, no. 19, 2015.

[7] D. Khanaferov, C. Luc, and T. Wang, "Social Network Data Mining using Natural Language Processing and Density based Clustering," in *Proc. ICSC*, California, 2014, pp. 250–251.

[8] P. Boersma, "Praat, A System for doing Phonetics by Computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[9] H. Kamper, F. D. Wet, T. Hain, and T. Niesler, "Capitalising on North American Speech Resources for the Development of a South African English Large Vocabulary Speech Recognition System," *Computer Speech and Language*, vol. 28, no. 6, pp. 1255–1268, 2014.

[10] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. ICSLP*, Denever-Colorado, 2002, pp. 901–904.

[11] L. Loots and T. Niesler, "Automatic Conversion between Pronunciations of Different English Accents," *Speech Communication*, vol. 53, pp. 75–84, 2010.

[12] K. Kintzley, A. Jansen, and H. Hermansky, "Featherweight Phonetic Keyword Search for Conversational Speech," in *Proc. ICASSP*, Florence, 2014, pp. 7859–7863.

[13] A. Mandal, K. R. P. Kumar, and P. Mitra, "Recent Developments in Spoken Term Detection: A Survey," *International Jour. of Speech Technology*, vol. 17, no. 2, pp. 183–198, 2014.

[14] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, R. C. Hsiao, G. Saikumar, I. Bulyko, L. Nguyen, J. Makhoul, F. Grezl, M. Hannemann, M. Karafiat, I. Szoke, K. Vesely, L. Lamel, and V. B. Le, "Score Normalization and System Combination for Improved Keyword Spotting," in *Proc. ASRU*, Olomouc, 2013, pp. 210–215.

[15] S. Tsakalidis, R. Hsiao, D. Karakos, T. Ng, S. Ranjan, G. Saikumar, L. Zhang, L. Nguyen, R. Schwartz, and J. Makhoul, "The 2013 BBN Vietnamese Telephone Speech Keyword Spotting System," in *Proc. ICASSP*, Florence, 2014, pp. 7829–7833.

[16] S. Wegmann, A. Faria, A. Janin, K. Reidhammer, and N. Morgan, "The Tao of ATWV: Probing the Mysteries of Keyword Search Performance," in *Proc. ASRU*, Olomouc, 2013, pp. 192–197.

[17] M. Akbacak, L. Burget, W. Wang, and J. van Hout, "Rich System Combination for Keyword Spotting in Noisy and Acoustically Heterogeneous Audio Streams," in *Proc. ICASSP*, Vancouver, 2002, pp. 8267–8271.

[18] A. J. K. Thambiratnam, *Acoustic Keyword Spotting in Speech with Applications to Data Mining*. Queensland Univerrsity of Technology, Brisbane: PhD Thesis, 2005.

[19] M. Larson and G. J. F. Jones, "Spoken Content Retrieval: A Survey of Techniques and Technologies," *Foundations and Trends in Information Retrieval*, vol. 5, no. 4-5, pp. 235–422, 2012.

[20] J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. London: Kluwer Academic Publishers, 1997.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, IEEE Catalog No: CFP11SRW-USB.

[22] M. Mohri, "Finite-State Transducers in Language and Speech Processing," *Association of Computational Linguistics*, vol. 23, no. 2, pp. 269–311, 1997.

[23] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafit, A. Rastrow, and R. C. Rose, "The Subspace Gaussian Mixture Model-A Structured Model for Speech Recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 204–439, 2011.

[24] D. Povey, X. Zhang, and S. Khudanpur, "Parallel Training of DNNs with Natural Gradient and Parameter Averaging," *arXiv preprint arXiv:1410.7455*, 2014.

[25] H. Jiang, "Confidence Measures for Speech Recognition: A Survey," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[26] Y. Wang and F. Metze, "An In-Depth Comparison of Keyword Specific Thresholding and Sum-to-One Score Normalization," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 2474–2478.

[27] B. Zhang, R. Schwartz, S. Tsakalidis, L. Nguyen, and S. Matsoukas, "White Listing and Score Normalization for Keyword Spotting of Noisy Speech," in *Proc. INTERSPEECH*, Portland, 2012, pp. 1832–1835.